



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Digital-linguistische Annotation
neulateinischer Texte.

Theorie und Praxis für Editions- und
Geschichtswissenschaft.

verfasst von / submitted by

Mag. Michael Frössl

angestrebter akademischer Grad / in partial fulfillment of the requirements for the
degree of

Master of Arts (MA)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

UA 066 804

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Masterstudium Geschichtsforschung,
Historische Hilfswissenschaften und Archivwissenschaft

Betreut von / Supervisor

PD Mag. Dr. Thomas J.J. Wallnig, MAS

Danksagung und Widmung

Methodisch betrachtet betritt der Verfasser vorliegender Studie mit dieser Arbeit größtenteils „Neuland“, verglichen mit seinen bisherigen Forschungsschwerpunkten. Es zählt hierbei zu den prägendsten Erfahrungen im Rahmen ihrer Erstellung, dass Wissenschaft seit jeher vom wertschätzenden Umgang, von Anerkennung und von gegenseitigen Anregungen und Austausch unter Kolleginnen und Kollegen lebt. Insbesondere für zeitgemäße Forschung unter dem Paradigma digital generierter bzw. digital unterstützter Ergebnisse scheint dies zu gelten: *Digital Humanities* gibt es nur im Plural, nur im Team. Insofern soll vorliegender Arbeit eine Danksagung vorangestellt werden, gerade weil ihr Verfasser im Rahmen ihrer Erstellung sehr viel Wertschätzung und Unterstützung von der Scientific Community erfahren hat – und das, obwohl vorliegende Arbeit weder höchsten Ansprüchen seitens der DH noch der Linguistik genügen wird, sondern allenfalls solchen digital interessierter Geschichtsforschung sowie jenen der Klassischen Philologie.

Allen voran sei meinem Betreuer Thomas Wallnig an dieser Stelle auf das Herzlichste gedankt für die grundsätzliche Idee zu dieser Arbeit, für seine großartige Unterstützung, seinen Rat und sein stets offenes Ohr während unserer vielen Gespräche, die den Arbeitsprozess begleiteten. Mit Thema und Ausrichtung der Arbeit hat er mich für die Wichtigkeit digitaler Forschung sensibilisiert und mir viele neue Perspektiven und Einblicke in diesen Bereich eröffnet. Ihm und allen Mitgliedern des Instituts für Österreichische Geschichtsforschung sei vorliegende Arbeit gewidmet. Von diesen sei insbesondere Manuela Mayer herzlich für ihre Anregungen gedankt. Ferner möchte ich den Kolleginnen und Kollegen am Austrian Centre for Digital Humanities der ÖAW in Wien Dank aussprechen: Claudia Resch, Daniel Schopper & Hannes Pirker, die mich in vielen technischen und inhaltlichen Belangen während des Arbeitsprozesses berieten und stets offen waren für Fragen meinerseits, die sich vielfach aus einem methodisch ganz traditionellen, nicht digitalen Verständnis von Forschung speisten. Außerdem möchte ich Martina Scholger und Sarah Lang vom Zentrum für Informationsmodellierung der Universität Graz für ihre Anregungen herzlich danken sowie Tim Geelhaar von der Goethe Universität Frankfurt am Main und Giuseppe G.A. Celano von der Universität Leipzig. Ohne die Scientific Community wäre die vorliegende Arbeit nicht zustande gekommen.

Wien, im Juli 2019

Digital-linguistische Annotation neulateinischer Texte.

Theorie und Praxis für Editions- und Geschichtswissenschaft

Inhaltsverzeichnis

1) Theorie und Präliminaria.....	1
1.1) Thematisch-methodische Hinführung.....	1
1.1.1) Ziele, Methoden und Inhalte vorliegender Arbeit.....	1
1.1.2) DH und Computerlinguistik?.....	9
1.1.3) Exemplarische Forschungsfragen.....	11
1.2) Terminologische Skizzen.....	12
1.2.1) Linguistik.....	12
1.2.2) Soziolinguistik.....	14
1.2.3) Historische Soziolinguistik.....	16
1.2.4) Computerlinguistik und Texttechnologie.....	19
1.2.5) Corpuslinguistik und sprachliche Corpora.....	21
1.2.6) Digital-linguistische Annotation und XML.....	28
1.2.6.1) Tokenisierung, Tagging, Lemmatisierung.....	31
1.2.6.2) PoS-Tagging und morphologische Annotation.....	33
1.2.6.3) Treebanking und Parsing.....	38
1.2.6.4) Rekapitulation.....	42
1.2.6.5) Semantisch-pragmatische Annotation.....	42
1.2.6.6) Soziolinguistische Annotation.....	43
1.2.6.7) Diskurs-/Textlinguistische Annotation.....	44
1.2.6.8) Fehlerannotation.....	44
1.2.6.9) Codierung und XML.....	45
2) Praktischer Hauptteil.....	52
2.1) Lateinische Tagsets für PoS und Morphologie.....	52
2.1.1) Allgemeine Vorbemerkungen.....	52
2.1.2) Tagging-Tools, Tagsets und Modelle.....	53
2.1.2.1) Das Morpheus Morphological Tagset.....	59
2.1.2.2) Das Tagset Lamap.....	63

2.1.2.3) Das PoS-Tagset der Sketch Engine.....	67
2.1.2.4) Das morphologische Tagset des Index Thomisticus Online.....	71
2.1.2.5) Das „Frankfurter Tagset“ und seine Erweiterungen.....	80
2.1.2.6) Das Tagset von LatMor.....	87
2.1.2.7) Das PoS-Tagset der Universal Dependencies.....	92
2.1.3) Zusammenfassung, Desiderate, Empfehlungen.....	95
2.2) Tokens, Kontext und Semantik.....	97
2.3) Das Werkzeug: Der TokenEditor.....	102
2.4) Zur Annotation der Eckhart-Briefe.....	108
2.4.1) Einzelprobleme und Schlussfolgerungen.....	109
2.4.2) Maschinelle Annotation, händische Eingriffe und Änderungen.....	115
2.5) Corpusabfragen und -auswertung.....	117
2.5.1) Forschungsfragen und quantitative Ergebnisse.....	118
2.5.2) Semantische Einzelanalysen.....	120
2.6) Schlussbetrachtung.....	131
3) Quellen- und Literaturverzeichnis.....	137
3.1) Gedruckte Quellen (Quelleneditionen).....	137
3.2) Sekundärliteratur.....	137
3.3) Wörterbücher und Grammatiken.....	153
3.4) Ausgewählte Online-Ressourcen.....	154
4) Anhang 1: Liste ausgewählter maschineller Annotationsfehler.....	157
5) Anhang 2/1: Belegstellen für semant. Analysen 1 (aus Briefen Eckharts an die Brüder Pez).....	159
6) Anhang 2/2: Belegstellen für semant. Analysen 2 (aus Briefen der Brüder Pez).....	169
7) Anhang 3: Abstract.....	183

1) Theorie und Präliminaria

1.1) Thematisch-methodische Hinführung

1.1.1) Ziele, Methoden und Inhalte vorliegender Arbeit

Dass die Geschichts-, die Editions- und die Literaturwissenschaften die Notwendigkeit erkannt haben, die Digitalisierung einerseits als historisches Phänomen zu untersuchen¹ und sie andererseits auch methodisch zu nutzen, davon legen zahlreiche Publikationen und Projekte der letzten Jahre auch im deutschsprachigen Raum beredtes Zeugnis ab.² „[F]ast durchgehend [haben die Editionswissenschaften sich in den letzten Jahren] auf digitale Methoden und Kommunikationsformen eingestellt.“³

Grundsätzlich geht es aus Sicht der vorliegenden Studie vor allem darum, nicht edierte, historische Texte („Quellen“) für Computer lesbar, für den Benutzer/die Benutzerin digital (also: in Form von digitalen Editionen) verfügbar zu machen und mit möglichst vielen zweckdienlichen Eigenschaften im virtuellen Kontext zu versehen, um Historikerinnen und Historikern die Arbeit mit elektronisch aufbereiteten Sprachdenkmälern zu erleichtern.⁴

Die vorliegende Studie versteht sich als gedrucktes Protokoll zu einem digitalen Pilotprojekt. Seine Ergebnisse bestehen im Wesentlichen aus zwei Teilen, erstens aus der hier vorliegenden Arbeit; zweitens wurden im Rahmen des Projektes

¹ Vgl. e.g. Thomas *Walach*, *Geschichte des virtuellen Denkens* (Wiesbaden 2018).

² Als Beispiel sie hier auf das Editionsprojekt rund um Berliner Intellektuelle des späten 18. und frühen 19. Jahrhunderts verwiesen: Anne *Baillet* (Hg.), *Briefe und Texte aus dem intellektuellen Berlin um 1800* (Berlin 2010–2015), online unter: <https://www.literatur.hu-berlin.de/de/berliner-intellektuelle-1800-1830> und <http://www.berliner-intellektuelle.eu/>; letzter Zugriff auf beide URLs: 30.05.2019.

³ Fotis *Jannidis*, *Perspektiven empirisch-quantitativer Methoden in der Literaturwissenschaft*. Ein Essay. In: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 89/4 (2015), 657–661; hier: 658.

⁴ Dass dabei von *nicht* nach modernen Kriterien edierten Quellen ausgegangen wird, ist innerhalb der Digital Humanities keineswegs selbstverständlich. Gerade ein Pionier- und quantitativ gewiss auch ein Mammut-Projekt der DH, die Perseus Digital Library, baut fast ausschließlich auf bereits edierten Texten auf, die entsprechend bearbeitet wurden, um sie in einer digitalen Umgebung verfügbar zu machen; vgl. Sarah *Lang*, *Review of Perseus Digital Library*. In: *Institut für Dokumentologie und Editorik e.V.* (Hg.), *ride. A Review Journal for Digital Editions and Resources* (Köln 2018); online unter: <https://ride.i-d-e.de/issues/issue-8/perseus/>; letzter Zugriff: 02.06.2019; eine Auflistung weiterer Publikationen zur Perseus Digital Library findet sich im weiteren Verlauf der vorliegenden Arbeit sowie im Literaturverzeichnis.

digitale Forschungsdaten generiert, die am Server des Austrian Centre for Digital Humanities der Österreichischen Akademie der Wissenschaften in Wien gespeichert sind. Interessierten Forscherinnen und Forschern können diese Daten auf Anfrage (auch zur weiteren Bearbeitung und Auswertung) zugänglich gemacht werden.

Ein Schwerpunkt des Projektes liegt zunächst in der untersuchten und angewandten Methodik: Es geht darum, sich ein computergestütztes Verfahren zu Eigen zu machen, zu diskutieren und umzusetzen, das gegenwärtig in den Bereichen der Computer-, der Corpuslinguistik, der Texttechnologie und des digitalen Editionswesens zur Anwendung gelangt. Das Potenzial dieser Methode innerhalb der Editionswissenschaften ist weitestgehend erwiesen, vor allem dort, wo sich in Abhängigkeit von den bearbeiteten Quellen und Forschungsfragen Forschungsinteressen der erwähnten Disziplinen mit denen verwandter wissenschaftlicher Bereiche treffen.⁵ Es handelt sich bei dieser Methode um sogenannte digital-linguistische Annotation.

Annotation bedeutet zunächst nichts anderes als die Anreicherung von vorhandener (Text-)Information mit zusätzlicher (additiver und/oder explizierend-erklärender) Information. Digitale, also computergestützte Annotation kann, muss aber nicht ausschließlich und nicht zwangsläufig nach linguistischen Kriterien erfolgen und kann als mögliche Vorstufe und als Vorarbeit für eine digitale Edition ausgewählter Quellen dienen. Was dies konkret bedeutet, ist in Kürze näher zu erläutern.

In Synergie mit digitalen Methoden bedient der Verfasser vorliegender Studie sich bei ihrer Erstellung auch traditioneller, nicht digitaler Arbeitsweisen. Die Auswertung der zunächst quantitativen, maschinell erstellten und in weiterer Folge manuell korrigierten Annotationsergebnisse erfolgt primär im Rahmen eines

⁵ Diese Disziplinen müssen nicht per se bzw. nicht ausschließlich im geisteswissenschaftlichen Bereich angesiedelt sein: Das Digitalisierungsprojekt rund um das Wien(n)erische Diarium, der Vorgängerzeitung der Wiener Zeitung, eine der ältesten bis heute erscheinenden Tageszeitungen weltweit, welches am Austrian Centre for Digital Humanities der ÖAW in Wien angesiedelt ist, wurde unter anderem für umwelthistorische Forschungen seitens der Wiener Universität für Bodenkultur genutzt. Dabei wurde innerhalb des Diariums nach auswertbaren Belegen für historische Hochwässer der Donau gesucht; vgl. <https://www.oeaw.ac.at/acdh/about/news-archive/news-detail/article/schlagzeilen-von-annodazumals-wandern-ins-web-1/>; allgemeiner zum Projekt einschließlich einer Liste bisheriger Publikationen: <https://www.oeaw.ac.at/acdh/projects/wienerisches-diarium-digital/>; letzter Zugriff auf beide URLs: 30.05.2019.

hermeneutisch-interpretativen Prozesses. Das annotierte Textcorpus dient hierfür als Ausgangsmaterial. Es stellt eine digitale Infrastruktur in kleinem Rahmen dar und ist über die vorliegende Studie hinaus für weitere Corpus-Abfragen offen. Dass Erstellung und selektive Auswertung der annotierten Texte in vorliegender Studie Hand in Hand gehen, ist keinesfalls selbstverständlich. Vielfach beschränken DH-Projekte sich (berechtigterweise – zumal auf Grund ihrer Größe) auf die *Bereitstellung* digitaler Infrastruktur. Abfragen, Auswertungen und Interpretation werden häufig erst nach deren Fertigstellung durchgeführt – fallweise von Forscherinnen und Forschern, die nicht am Erstellungsprozess beteiligt waren.

Im Rahmen der vorliegenden Studie wurden ausgewählte frühneuzeitliche Gelehrtenbriefe (meist) neulateinischer Sprache linguistisch annotiert, jene des ursprünglich protestantischen, später katholischen Historikers und Bibliothekars Johann Georg Eckhart (gest. 1730 in Würzburg) an die Geschichtsforscher Bernhard und Hieronymus Pez OSB (gest. 1735 bzw. 1762 in Melk).

Im Zentrum der Arbeit stand dabei die Frage, welche Standards und welche Varianten linguistischer Annotation in Hinblick auf die digitale Edition speziell neulateinischer Quellen der frühen Neuzeit am besten geeignet erscheinen.⁶

⁶ Die Notwendigkeit des hier skizzierten Arbeitsschwerpunktes ergab sich im Laufe der Erstellung vorliegender Studie. Deren Zielsetzung erfuhr dadurch eine wesentliche Veränderung im Vergleich zu ihrer ursprünglichen Intention. Anfangs war geplant gewesen, größere Teile der Pez-Korrespondenz linguistisch zu annotieren, um auf dieser Basis sprachwissenschaftliche Vergleiche zwischen mehreren Korrespondenzpartnern unterschiedlicher Muttersprache (u.a. Deutsch, Italienisch, Französisch) durchzuführen. Von diesem Vorhaben musste im Wesentlichen aus zwei Gründen Abstand genommen werden: erstens wegen des beträchtlichen Aufwandes, der mit linguistischer Annotation eines historischen Textcorpus verbunden ist – aus diesem Grund musste auf die Annotation mehrerer Parallelcorpora verzichtet werden; zweitens, weil bestehende lateinische Textcorpora, die über linguistische Annotationen verfügen, bisher unter Anwendung höchst heterogener Richtlinien annotiert wurden. Diese Richtlinien (Tagsets) bedurften zunächst einer vergleichenden Evaluation. Insofern stellen sich jene Forschungsergebnisse, die im Rahmen vorliegender Studie erzielt wurden, als etwas anderes dar als ursprünglich geplant.

Corpusgestützte digitale Suchabfragen mit dem Ziel, die Sprachgewohnheiten mehrerer lateinisch schreibender Gelehrter unterschiedlicher Muttersprache zu vergleichen, mussten im Lichte dessen, dass im Wesentlichen bloß *ein* Textcorpus maschinell annotiert und händisch nachkorrigiert werden konnte, auf ein Minimum beschränkt bleiben. Insofern es sich bei den erzielten Forschungsergebnissen im einfachsten Fall um bloße Wortfrequenzlisten handelt, mögen diese Forschungsergebnisse durchaus bescheiden anmuten. Immerhin konnten sie als Ausgangsbasis für einfache semantische Untersuchungen und Vergleiche dienen. Erschwerend kam hinzu, dass das annotierte Textcorpus – die Briefe Eckharts an Pez – neben Latein auch deutschsprachige Passagen enthält, was bei der automatischen Vorannotation zu Problemen für die Annotationssoftware führte.

Zur Diskussion standen einerseits linguistische Basisannotation mit Kennzeichnung der Wortarten und mit Ausweis der Grundworte (Lemmatisierung), andererseits morphologische Annotation und syntaktisches Treebanking. Erklärungen und Details zu diesen Begriffen sind im weiteren Verlauf der Arbeit zu finden.

Hierbei wird auf einer intensiven theoretischen Auseinandersetzung mit dem aktuellen Forschungsstand rund um linguistische Annotation aufgebaut. Vorliegende Arbeit bietet zu diesem Thema einen kompakten Überblick, indem sie aus bestehender Literatur schöpft. Ferner wurden Annotationsschemata (Tagsets) evaluiert, die im Rahmen computerlinguistischer bzw. computerphilologischer Forschungsprojekte der jüngeren Vergangenheit auf Texte lateinischer Sprache angewandt wurden (v.a. im Rahmen der Perseus Digital Library sowie des Index Thomisticus Online⁷).⁸

Wegen ihres fallweise beträchtlichen Aufwandes wird digital-linguistische Annotation im Rahmen von editionswissenschaftlichen Projekten nicht immer angewandt. Dennoch kann linguistische Annotation als eine weitestgehend anerkannte Methode der Digital Humanities bzw. der Corpuslinguistik angesehen werden. Zumeist wird sie auf modernsprachliche Textsammlungen angewandt. Linguistisch annotierte Corpora älterer Texte sind im Vergleich dazu relativ selten, zumal in der deutschsprachigen Corpuslandschaft.⁹ Gleichwohl wurden bzw. werden gerade in jüngster Zeit große Bemühungen unternommen, um den relativen Mangel an historischen Textcorpora auszugleichen (z.B. Deutsches Textarchiv, Referenzkorpus Mittelhochdeutsch et c.).

⁷ Ausgewählte Beiträge zu verschiedenen linguistischen Annotationsarten, zur Perseus Digital Library sowie zum Index Thomisticus Online finden sich im Literaturverzeichnis.

⁸ Neulateinische Texte übertreffen jene der Antike und des Mittelalters quantitativ zwar bei Weitem, doch sind sie, zumal Gelehrtenkorrespondenzen, in der computerlinguistischen Forschung sowie im Bereich der Digital Humanities bisher eher unterrepräsentiert. Eine Ausnahme stellt in diesem Zusammenhang das ePistolarium dar; vgl. <http://ckcc.huylgens.knaw.nl/>; letzter Zugriff: 30.05.2019.

⁹ Vgl. Ulrike Czeitschner, Claudia Resch, Morphosyntaktische Annotation historischer deutscher Texte: Das Austrian Baroque Corpus. In: Wolfgang Dressler, Claudia Resch (Hgg.), Digitale Methoden der Korpusforschung in Österreich (Veröffentlichungen zur Linguistik und Kommunikationsforschung, Bd. 30, Wien 2017), 39–62, insbes. 40 f. (einschließlich dortiger Anm. 7 sowie Anm. 24).

Durch linguistische Annotation wird eine Ausgangsbasis geschaffen für die (vergleichende) sprachwissenschaftliche Auswertung mitunter sehr großer Textquanten. Die in einem Text implizit enthaltenen linguistischen Informationen (z.B. Wortarten und Morphologie) werden durch linguistische Annotation explizit gemacht, maschinell les- und verarbeitbar.

Ein besonderer Nutzen annotierter Corpora ist „insbesondere [in] der Identifikation von musterhaften Regularitäten, Häufungen und Wiederholungserscheinungen“¹⁰ eines Textes zu sehen, was dem Prinzip nach sprachübergreifend gelten kann.

Eine terminologische Differenzierung zwischen digitalen Methoden einerseits und digitalen Werkzeugen andererseits ist insofern unerlässlich, als die Methode der Annotation zwar auch manuell, ohne Unterstützung durch den Computer, vorgenommen werden kann. So sie aber digital erfolgen soll, ist sie notwendigerweise auf spezielle Werkzeuge der elektronischen Datenverarbeitung (Rechner, Tools, Apps bzw. Programme) angewiesen, im Rahmen derer sie an ausgewählten Texten vollzogen wird, die ihrerseits digital codiert und daher für den Computer lesbar vorliegen.

Eine wesentliche Eigenschaft digital annotierter Editionen stellt in der Regel ihre gezielte Durchsuchbarkeit mittels computergestützter Suchabfrage dar, entweder durch einfache Eingaben im Suchfeld oder durch komplexere Suchanfragen mittels formaler Abfragesprachen wie z.B. mittels spezifischer CQL – Contextual Query Language – oder mittels XPath. Eine gedruckte Edition ist nur dann durchsuchbar, wenn sie mit einem Register, etwa für Personen- oder Ortsnamen versehen wurde. Ähnliches gilt *mutatis mutandis* für digitale Editionen: Möglichst umfassend annotierte Texte werden zum digitalen Register, ja zur vollständigen Konkordanz ihrer selbst in sich selbst. Ihrer gezielten Durchsuchbarkeit geht bei ihrer Erstellung notwendigerweise die digitale Annotation als *condicio sine qua non* voraus, wobei Letztere bei Weitem nicht in allen, aber in bestimmten Fällen (und mit jeweils unterschiedlich genauen Ergebnissen) vollautomatisch (nur durch den Computer) oder halbautomatisch (einschließlich menschlicher Nachkontrolle der computergenerierten Ergebnisse) erfolgen kann oder ausschließlich manuell.

¹⁰ Czeitschner, Resch, Austrian Baroque Corpus (wie Anm. 9), 52.

Scannt man dagegen einen gedruckt oder handschriftlich vorliegenden Text einfach nur ein, so ist digitale Durchsuchbarkeit bekanntlich noch lange nicht gegeben.

Begrenzt sind die Möglichkeiten systematischer Durchsuchbarkeit auch im Falle von Texten, die (X)HTML-codiert zwar im World-Wide-Web vorliegen, nicht aber (linguistisch) annotiert sind. Die Suchfunktion von Web-Browsern oder gängigen Schreibprogrammen (Strg+F) generiert zumeist nämlich nur solche Ergebnisse, die exakt der Zeichenfolge der Eingabe entsprechen. Durchsucht man einen (X)HTML-codierten Text im World-Wide-Web beispielsweise nach der Zeichenfolge *Heirat*, so werden nur Ergebnisse angezeigt, die exakt dieser Suchvorgabe entsprechen. Orthographische Varianten des Grundwortes *Heirat*, etwa in historischer Schreibweise *Heyrath* oder *Heurath* werden nicht angezeigt. Führt man dieselbe Suche jedoch innerhalb eines linguistisch annotierten Textes durch, so erscheinen auch alle im Text vorhandenen, abgewandelten Formen des Grundwortes und alle historischen Schreibweisen, ohne dass man sie einzeln in die Suchmaske eingeben müsste. Voraussetzung dafür ist die digitale Anreicherung einer jeden Zeichenkette mit ihrem Lemma, dem Grundwort. Dieser Vorgang ist bei digital-linguistischer Annotation von größter Bedeutung und übertrifft andere Teilschritte der Annotierung an Stellenwert: „Ein Lemma [...] steht stellvertretend für das gesamte Flexionsparadigma und führt nicht nur zu allen im Korpus vorkommenden Wortformen [...], sondern auch zu allen Schreibvarianten.“¹¹

Um ein Pilotprojekt handelt es sich im vorliegenden Fall deshalb, weil die Methode der digital-linguistischen Annotation auf eine Quellengattung angewandt wird, bei der dies im deutschen Sprachraum bisher noch nicht geschehen ist.¹²

Bei der ausgewählten Quellengattung handelt es sich, wie bereits kurz angedeutet, um den neulateinischen Gelehrtenbrief des frühen 18. Jahrhunderts.¹³ Im Fokus

¹¹ Czeitschner, Resch, Austrian Baroque Corpus (wie Anm. 9), 47.

¹² Dem Verfasser der vorliegenden Zeilen sind gegenwärtig keine vergleichbaren Projekte im deutschsprachigen Raum bekannt. Auf das ePistolarium wurde bereits hingewiesen (vgl. Anm. 8). Die Methode der digitalen Annotation und die ausgewählte Quellengattung bedingen einander nicht. Vielmehr verdankt sich ihr Zusammentreffen nicht zuletzt zufälligen Begebenheiten.

¹³ Vgl. e.g. Thomas Wallnig, Gelehrtenkorrespondenzen und Gelehrtenbriefe. In: Josef Pauser, Martin Scheutz, Thomas Winkelbauer (Hgg.), Quellenkunde der Habsburgermonarchie (16.–18. Jahrhundert). Ein exemplarisches Handbuch (Mitteilungen des Instituts für Österreichische Geschichtsforschung, Ergänzungsband 44, Wien/München 2004), 813–827.

der Anwendung steht eine ausgewählte Teilkorrespondenz aus dem umfangreichen gelehrten Schriftverkehr der Brüder Bernhard und Hieronymus Pez OSB, die als benediktinische Ordens- und Landes(literar)historiker von Stift Melk im heutigen Niederösterreich aus tätig waren und die durch ihr Wirken wesentlich zur Entwicklung des historisch-kritischen Methodenspektrums im katholischen Süden des Heiligen Römischen Reiches, im süddeutsch-bayrischen Raum, beitrugen.¹⁴

Nach festgelegten Kriterien sowie unter Weiterentwicklung bestehender Annotationsstandards und unter Verwendung von Werkzeugen, die das Austrian Centre for Digital Humanities (ACDH) der ÖAW in Wien zur Verfügung stellte, wurden die erhaltenen neulateinischen Briefe Johann Georg Eckharts, des vormaligen Sekretärs von Gottfried Wilhelm Leibniz, an Bernhard Pez (entstanden in der Periode zwischen 1717 und 1728) linguistisch annotiert (Wortarten und Morphologie).¹⁵

Die Annotation erfolgte zunächst computergestützt (automatisch) und umfasste alle sprachlichen Bestandteile (Tokens) der genannten Quellen, sofern es sich nicht um deutschsprachige Passagen handelte, die vereinzelt auftreten. Neben dem Münchner Tagging-Tool MarMot¹⁶ wurde dabei der sogenannte TokenEditor verwendet, eine Web-Applikation entwickelt am ACDH in Wien.¹⁷ Die auf diesem Wege generierten Annotationsergebnisse wurden händisch korrigiert – einschließlich Morphologie von rund 14.850 sprachlichen Einheiten. Erfahrungen,

¹⁴ Ausgewählte Sekundärliteratur hierzu im Literaturverzeichnis.

¹⁵ Die erhaltenen Briefe Johann Georg Eckharts an Bernhard Pez liegen gegenwärtig (Stand Juni 2019) in folgender Ausgabe gedruckt vor: Thomas Wallnig, Die Briefe von Johann Georg Eckhart an Bernhard Pez in Text und Kommentar (Staatsprüfungsarbeit am Institut für Österreichische Geschichtsforschung, Wien 2001).

Bis einschließlich 1718 wurden sie auch ediert im Rahmen der bisherigen Gesamtedition der Pez-Korrespondenz: Thomas Wallnig, Thomas Stockinger, Die gelehrte Korrespondenz der Brüder Pez. Text, Regesten, Kommentare, Bd. 1: 1709–1715 (Quelleneditionen des Instituts für Österreichische Geschichtsforschung, Bd. 2/1, München/Wien 2010) und:

Thomas Stockinger, Thomas Wallnig, Patrick Fiska, Ines Peper, Manuela Mayer, Die gelehrte Korrespondenz der Brüder Pez. Text, Regesten, Kommentare, Bd. 2 (2 Halbbände): 1716–1718 (Quelleneditionen des Instituts für Österreichische Geschichtsforschung, Bd. 2/1 und Bd. 2/2, Wien 2015).

¹⁶ Dieser Tagger wurde ursprünglich an Hand von arabischen, tschechischen, spanischen, deutschen und ungarischen Texten getestet; vgl. Thomas Müller, Helmut Schmid, Hinrich Schütze, Efficient Higher-Order CRFs for Morphological Tagging. In: *Association for Computational Linguistics* (Hg.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle 2013), 322–332; online unter: <http://www.aclweb.org/anthology/D13-1032>; Web-Präsenz des Taggers: <http://cistern.cis.lmu.de/marmot/>; letzter Zugriff auf beide URLs: 30.05.2019.

¹⁷ <https://www.oew.ac.at/acdh/tools/tokeneditor/>; letzter Zugriff: 30.05.2019; vgl. auch Czeitschner, Resch, Austrian Baroque Corpus (wie Anm. 9), 48 f.

die sich im Zuge der praktischen Annotation ergaben, sollen im Rahmen der vorliegenden Arbeit schriftlich zusammengefasst und die erwähnten linguistischen Annotationsmethoden kontrastiert und verglichen werden (morphologische Annotation und linguistische Basisannotation mit Kennzeichnung der Wortarten und Lemmatisierung gegenüber syntaktischem Treebanking; siehe unten).

Von fast allen Briefen, die annotiert wurden, existierten zum Zeitpunkt der Annotation bereits historisch-kritische Editionen (vgl. Anm. 15). Der Aspekt der Quellenbearbeitung floss im Rahmen des vorliegenden Projektes insofern ein, als die edierten Briefe Eckharts an Bernhard Pez noch einmal einer Kollationierung mit den digitalisierten handschriftlichen Originalen unterzogen wurden,¹⁸ bevor ihr um editorische Anmerkungen bereinigter Rohtext in das erwähnte Annotationswerkzeug, den TokenEditor, eingespielt wurde.¹⁹ Letzteres geschah in enger Zusammenarbeit mit dem ACDH der ÖAW in Wien.²⁰

Die Anwendbarkeit der digital-linguistischen Annotation ist dabei weder auf die Quellengattung des Gelehrtenbriefes beschränkt, noch spielt es eine Rolle, ob es sich um Amts- bzw. um Geschäftsschriftgut, um semiliterarische oder literarische Texte im strengen Sinn handelt, die digital annotiert werden. Grundsätzlich kann *jeder* Text digital annotiert werden, sofern er nur der digitalen Bearbeitung zugänglich ist und die entsprechende elektronische Infrastruktur vorliegt.

Die vorliegende Studie richtet sich vor allem an Forscherinnen und Forscher sowie an Angehörige einer breiteren interessierten Leserschaft, denen die computergestützte Arbeitsweise speziell zwecks der Generierung neuer Forschungsergebnisse bisher eher fremd war. Ihrer Begrifflichkeit nach ist die vorliegende Studie daher bewusst einfach gehalten. Sie möchte den Versuch unternehmen, Möglichkeiten der digital-linguistischen Annotation speziell im Bereich des neulateinischen Gelehrtenbriefes aus dem frühen 18. Jahrhundert auszuloten und zu diskutieren.

Da es sich bei digitaler Annotation um eine Methode der Computer- bzw. der Corpuslinguistik handelt, ergibt sich zunächst die Notwendigkeit, die Begriffe

¹⁸ Die Scans der handschriftlichen Briefe sind (mit Stand Mai 2019) einsehbar unter: <https://unidam.univie.ac.at/nachlass/195> (Suchmaske „Person“ > Eingabe: „Eckhart“ > Pez-Korrespondenz, Band II).

¹⁹ Für seine Unterstützung und für die Kontrolle bei der Kollationierung dankt der Verfasser dieser Zeilen Thomas Wallnig sehr herzlich.

²⁰ An der Durchführung dieser Prozesse waren Daniel Schopper und Hannes Pirker vom ACDH wesentlich beteiligt. Auch ihnen gilt an dieser Stelle herzlicher Dank.

Computer-, Corpuslinguistik sowie digitale Annotation einführend zu erläutern. Auf eine umfassende Darstellung formaler Grundlagen der Computerlinguistik sowie auf eine umfassende Darstellung computerlinguistischer Methoden muss in vorliegender Arbeit aus Platzgründen jedoch weitestgehend verzichtet werden.²¹ Vor allem auf digital-linguistische Annotation soll in der vorliegenden Arbeit fokussiert werden. Darüber hinaus sollen einige Grundbegriffe der Linguistik näherhin erläutert werden, sofern diese für ein grundlegendes Verständnis der digitalen Annotation zweckdienlich erscheinen. Ein weiteres Hauptanliegen der vorliegenden Arbeit besteht, wie bereits kurz erwähnt, in der vergleichenden Evaluation jener Annotationsinventare (Tagsets), die in bisherigen Projekten des lateinischen Natural Language Processing bzw. der lateinischen DH verwendet wurden.

1.1.2) DH und Computerlinguistik?

An erster Stelle lässt sich die Frage nach dem Verhältnis zwischen diversen digital arbeitenden Teilbereichen der Linguistik einerseits und den Digital Humanities bzw. ihren Vorgängerdisziplinen andererseits aufwerfen, ganz allgemein sowie in geschichtlicher und methodischer Hinsicht. Es ist im Rahmen der vorliegenden Arbeit zwar nicht möglich, diese Fragen auch nur annähernd umfassend zu beantworten, doch sei – gleichsam stichwortartig – auf folgende Punkte hingewiesen: Ab den 1970er- und ab den 1980er-Jahren vollzog sich nicht zuletzt durch die akademische Etablierung der Computerlinguistik und der Texttechnologie eine „deutliche Trennung zwischen dem Computereinsatz in den Sprachwissenschaften und allen anderen Geisteswissenschaften, einschließlich der literaturwissenschaftlichen Disziplinen [...]“.²²

Zwar scheint es gegenwärtig so, als würden die Digital Humanities für sich beanspruchen, auch die (Computer-)Linguistik zu umfassen.²³ Praktisch gesehen allerdings adressieren die DH doch in erster Linie Vertreter/-innen der

²¹ Hierzu sei beispielsweise auf das enzyklopädische Nachschlagewerk von Carstensen (et al.), das in dritter Auflage im Jahre 2010 erschienen ist, verwiesen: Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde, Hagen Langer (Hgg.), Computerlinguistik und Sprachtechnologie. Eine Einführung (Heidelberg ³2010).

²² Manfred Thaller, Geschichte der Digital Humanities. In: Fotis Jannidis, Hubertus Kohle, Malte Rehbein (Hgg.), Digital Humanities. Eine Einführung (Stuttgart 2017), 3–12; hier: 7.

²³ Vgl. Manfred Thaller, Digital Humanities als Wissenschaft. In: Jannidis, Kohle, Rehbein, Digital Humanities (wie Anm. 22), 13–18; hier: 13.

Editionswissenschaften, der Philologien, der Geschichtswissenschaften sowie ihrer Sub- und Nachbardisziplinen, die unter dem Dach der historischen Kulturwissenschaften „durch die Dimension der geschichtlichen Betrachtungs- und Vorgehensweise eng miteinander verbunden sind“²⁴, das heißt: z.B. der Archäologie, der (Kultur- und der Sozial-)Anthropologie, der Ethnologie usw.; die DH verstehen sich demnach als digital arbeitende (historische) Kulturwissenschaften, als ein „Komplex kulturwissenschaftlicher Fächer“²⁵, bei denen – im Grunde genommen – „die Aufspaltung der historischen Dimension auf Dutzende von Fachdisziplinen [...] nicht erforderlich [...], nicht erwünscht ist.“²⁶ Zu den verschiedensten Teildisziplinen der Linguistik aber scheint seit den 1970er- und den 1980er-Jahren nach wie vor kein allzu großes Naheverhältnis mehr zu bestehen (etwa hinsichtlich gemeinsamer Fragestellungen, gemeinsamer Herangehensweisen oder hinsichtlich gemeinsamer Projekte).²⁷

Einer der Gründerväter der Digital Humanities, Roberto Busa (1913–2011), wird, soweit ich derzeit sehe, seitens der deutschsprachigen Computerlinguistik nicht flächendeckend als solcher für die eigene Disziplin reklamiert²⁸ und noch 2017 konnten Mitglieder der TEI Special Interest Group for Linguists (LingSIG) konstatieren, es herrsche „[a] lack of a specific [standard] for lightweight linguistic annotation within the TEI“²⁹, während bestehende Möglichkeiten für TEI-konforme linguistische Annotation unverhältnismäßig kompliziert seien. Die Lücke zwischen DH und Computerlinguistik scheint sich erst seit kurzem und nur langsam wieder zu verkleinern, etwa durch die Etablierung einer neuen Attributklasse `@linguistic` in die TEI-Guidelines von Jänner 2018. Gleichwohl

²⁴ Wolfgang *Schmale*, Einleitung: Digital Humanities. Historische Kulturwissenschaften. In: Wolfgang *Schmale* (Hg.), *Digital Humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität* (Stuttgart 2015), 9 – 13; hier: 9.

²⁵ *Schmale*, *Dig. Humanities* (wie Anm. 24), 9.

²⁶ *Schmale*, *Dig. Humanities* (wie Anm. 24), 10.

²⁷ Verschiedene methodische Ansätze und Absichten scheinen auch in der terminologischen Unterscheidung zwischen Digital Humanities und Computational Humanities zu Tage zu treten, worauf in der vorliegenden Arbeit jedoch nicht näher eingegangen werden soll.

²⁸ In den wenigen von mir konsultierten deutschsprachigen Einführungen in die Computerlinguistik findet Busa keine Erwähnung; vgl. neben *Carstensen, Ebert et al.*, *Computerlinguistik und Sprachtechnologie* (wie Anm. 21), 18–25 auch: Henning *Lobin*, *Computerlinguistik und Texttechnologie* (Paderborn 2010), 14–17.

²⁹ Piotr *Bański*, Susanne *Haaf*, Martin *Mueller*, Lightweight Grammatical Annotation in the TEI. New Perspectives. In: *European Language Resource Association* (Hg.), [Proceedings of the Eleventh International Conference on Language Resources and Evaluation \(LREC-2018\)](http://www.lrec-conf.org/proceedings/lrec2018/pdf/422.pdf) (Miyazaki 2018), 1795–1802; hier: 1795; online unter: <http://www.lrec-conf.org/proceedings/lrec2018/pdf/422.pdf>; gesamter Tagungsband: <https://www.aclweb.org/anthology/L18-1> (letzter Zugriff auf beide URLs: 30.05.2019).

bauen die DH auch auf Methoden der Computer- bzw. der digitalen Corpuslinguistik sowie der Texttechnologie auf, allein schon durch die Anwendung der digital-linguistischen Annotation. Hier bestehen zahlreiche Überschneidungen. Insofern erschien es gerechtfertigt, im theoretischen Teil der vorliegenden Arbeit digitale Annotation nicht nur, aber auch (und vor allem) aus Sicht der Corpus- bzw. der Computerlinguistik zu erläutern.

Dieser Zugang hat sich insofern als glücklich erwiesen, als er für einen Einstieg in die Materie ausgesprochen geeignet erscheint. Einige der besten, weil einfach und verständlich geschriebenen Einführungen in den Bereich der digitalen Annotation kommen aus dem Bereich der Corpuslinguistik.³⁰

1.1.3) Exemplarische Forschungsfragen

Wo mehrere einheitlich annotierte Textcorpora derselben (oder: einer ähnlichen) Sprachstufe vorliegen (und zumal auch derselben Gattung bzw. desselben Schriftgut-Typs), können anhand der zunächst quantitativen Annotations-ergebnisse diachrone sprachlich-historische Vergleiche mittels digitaler Werkzeuge durchgeführt werden: Gibt es zwischen den ausgewählten Textcorpora sprachlich-stilistische Präferenzen und Unterschiede, die sich im Rahmen der angewandten Methoden eindeutig quantifizieren lassen? Welcher Text neigt mehr zur Hypotaxe, welcher mehr zur Parataxe (z.B. durch Suchabfrage nach der Anzahl unterordnender Bindeworte)? Welche Nebensätze treten wann und in welchem Text gehäuft auf? Gibt es lexikalische Vorlieben? Lassen sich Präferenzen in Richtung Verbal- oder Nominalstil mit digitalen Suchabfragen eindeutig feststellen? Diese und ähnlich geartete Fragen sollen ihrem Prinzip nach das Potential andeuten, welches linguistischer Annotation innewohnt sowie mehreren annotierten Parallelcorpora.

Mit den erwähnten methodischen Reflexionen zur linguistischen Annotation sowie mit der praktischen Annotation hauptsächlich *eines* ausgewählten Textcorpus kann die vorliegende Arbeit hierfür jedoch nur eine erste Ausgangsbasis bieten.

³⁰ Vgl. e.g. Carmen Scherer, *Korpuslinguistik* (Heidelberg 2006).

Ferner muss an dieser Stelle darauf hingewiesen werden, dass das Ergebnis der praktischen Annotierung, das annotierte Corpus der Eckhart-Briefe, mit Stand August 2019 noch nicht in einer digitalen Codierung vorliegt, die den gegenwärtigen Empfehlungen der TEI entspräche, also ohne die aktuelle Attribut-Gruppe „@linguistic“ (vgl. oben, insbesondere Anm. 29). Eine entsprechende Konversion ist in weiterer Folge geplant.

1.2) Terminologische Skizzen

1.2.1) Linguistik

Die Begriffe Computerlinguistik und Corpuslinguistik bezeichnen grundsätzlich zwei verwandte Teildisziplinen der Sprachwissenschaft, der Linguistik.³¹ Der Gegenstand der Linguistik ist die menschliche Sprache, die Tätigkeit des Linguisten/der Linguistin deren Untersuchung und Beschreibung. Unter den unscharfen, mehrdeutigen und daher vorwissenschaftlichen Begriff Sprache fallen im Sinne der Linguistik grob gesprochen:³² (1) Sprache als allgemein-menschliche Fähigkeit zur Kommunikation zwecks Erreichung verschiedenster Ziele (*langage*); (2) Sprache in ihrem Aufbau als System, das Merkmale aufweist, die allen menschlichen Einzelsprachen gemeinsam sind (sprachliche Universalien); (3) Sprache als Einzelsprache (*langue*), d.h. als spezielles menschliches Kommunikationssystem einer bestimmten Sprachgemeinschaft – die Gesamtheit der Ausdrucksmittel einer Einzelsprache; (4) Sprache als einzelner Sprechakt, als „das Sprechen“ – die Art und Weise, wie im konkreten sprachlichen Akt von den Ressourcen einer Einzelsprache Gebrauch gemacht wird (*parole*).

³¹ Fallweise wird im deutschen Sprachraum zwischen Sprachwissenschaft einerseits und Linguistik andererseits differenziert; vgl. Heidrun Pelz, Linguistik. Eine Einführung (Hamburg ⁴1999), 23f.; Gerhard Nickel, Einführung in die Linguistik. Entwicklung, Probleme, Methoden (Berlin 1979), 17; allerdings werden die beiden Begriffe einander auch gleichgesetzt; vgl. Heinz Vater, Einführung in die Sprachwissenschaft (München ²1994), 12, Anm. 3. Im vorliegenden Fall unterscheiden wir mit Vater und Nickel die Begriffe Linguist(ik) und Sprachwissenschaft(ler*in) nicht voneinander und betrachten sie als synonym.

³² Vgl. Pelz, Linguistik. Einführung (wie Anm. 31), 23f.; Vater, Einf. Sprachwissenschaft (wie Anm. 31), 11–14.

Zu den grundlegenden Teilbereichen, die das Fach Sprachwissenschaft strukturieren, zählen:³³ (1) Lautlehre (bestehend aus Phonetik und Phonologie); diese untersucht die Art der Sprachlaute, deren materiell bedingte Entstehungsweise, deren materielle Übertragungs- und Wahrnehmungsgrundlagen und deren physikalische Eigenschaften einerseits (Phonetik), andererseits das Vorkommen und die Funktionen der Laute in einzelnen Sprachen sowie die Regularitäten bei deren Verknüpfung (Phonologie); (2) Formenlehre (Morphologie), die die Wortstruktur analysiert: Wie werden kleinste bedeutungstragende Bausteine (Morpheme) zu Worten zusammengesetzt? Wie verhalten einzelne Morpheme sich zueinander, welche Funktionen kommen ihnen zu? (3) Satzlehre (Syntax): Sie beschäftigt sich mit der Frage, wie, nach welchen Regeln Worte „zu größeren strukturellen Einheiten (Wortgruppen, Sätzen) zusammengefügt werden und welche Funktion einzelne Teile des Satzes haben“³⁴; (4) Semantik: Sie fokussiert auf die Bedeutung von Worten sowie darauf, wie Bedeutungen in Phrasen oder Sätzen zusammenwirken; (5) Pragmatik: Sie fragt, wie umgebende Faktoren (Handlungen, Situationen) mit der Bedeutung sprachlicher Äußerungen zusammenhängen, wie sie diese beeinflussen.

Lautlehre, Morphologie und Syntax zählen zu den unbestrittenen Kernbereichen der Linguistik. Mit diesen Gebieten beschäftigt sich neben der Linguistik keine andere wissenschaftliche Disziplin in ähnlich umfassender, profunder und systematischer Art und Weise. Vorbehaltlich geringer Einschränkungen gilt dies auch für die Semantik, wohingegen die Pragmatik, die Überschneidungen mit letzterer aufweist, zwar von vielen, nicht aber von allen Linguistinnen und Linguisten zu den Kernbereichen der Sprachwissenschaft gezählt wird.³⁵ Aus Sicht der Geschichtswissenschaft, der Philologien und aus Sicht der Digital Humanities insgesamt scheinen jedoch gerade diese zwei zuletzt genannten Teilbereiche, Semantik und Pragmatik, von besonderem Interesse zu sein, speziell, wenn es darum geht, bei Fragen der Mentalitäts- und der Kulturgeschichte nach den

³³ Vgl. Peter *Schlobinski*, Grundfragen der Sprachwissenschaft. Eine Einführung in die Welt der Sprache(n) (Göttingen 2014), 15; *Vater*, Einf. Sprachwissenschaft (wie Anm. 31), 24.

³⁴ *Schlobinski*, Grundfragen Sprachwissenschaft (wie Anm. 33), 15.

³⁵ Hinsichtlich Semantik weist *Vater* darauf hin, dass diese lange Zeit Gegenstand der Philosophie war, während in jüngerer Zeit zwischen philosophischer und linguistischer Semantik unterschieden wird; vgl. *Vater*, Einf. Sprachwissenschaft (wie Anm. 31), 24f. (einschließlich dortiger Anmerkung 12) & 144f.

gedanklichen Konzepten „hinter“ einzelnen Begriffen zu suchen: Roberto Busa fragte in seiner Dissertation nach dem Konzept hinter dem Begriff der *praesentia* bei Thomas von Aquin. Sie trägt den Titel: *La Terminologia Tomistica dell'interiorità. Saggi di metodo per un' interpretazione della metafisica della presenza*.

In Summe ist das Phänomen Sprache so komplex, dass ihre umfassende Bearbeitung „nicht von der Sprachwissenschaft alleine geleistet werden [kann], sondern [nur] in Zusammenarbeit mit Nachbardisziplinen wie Psychologie, Philosophie, Soziologie, Medizin und Naturwissenschaften allgemein.“³⁶ Je nach dem, mit welcher Nachbardisziplin ein Linguist/eine Linguistin zusammenarbeitet bzw. mit welchem Schwerpunkt er/sie die eigenen Forschungen betreibt, ergeben sich unterschiedlich akzentuierte Spezial- respektive Mischformen von Linguistik, z.B. Bio-, Computer-, Corpus-, Diskurs-, Gender-, Kontakt-, Psycho-, Sozio- oder Textlinguistik usw.; dem ironisch gefärbten Begriff „Bindestrich-Linguistik“ scheint mittlerweile topischer Charakter zuzukommen. Neben der Computer- und der Corpuslinguistik, auf die wir in Kürze näher eingehen werden, ist für uns im Kontext der Geschichtswissenschaft besonders die historische Soziolinguistik (*historical sociolinguistics* oder *socio-historical linguistics*) von Interesse. Diese Forschungsdisziplin ist relativ jung. Sie baut auf den Grundfragen der Soziolinguistik auf.

1.2.2) Soziolinguistik

Soziolinguistik³⁷, entstanden in den USA während der 1950er- und der 1960er-Jahre, beschäftigt sich mit den Verschränkungen zwischen Gesellschaft und Sprache. Sie untersucht das wechselseitige Bedingungsgefüge, das zwischen Sozial- und Sprachstruktur(en) besteht. Ihre Aufgabe ist es, die gesellschaftlichen Voraussetzungen, Implikationen, Funktionen und Bewertungen von Sprachsystemen und Sprachgebrauch, vor allem innerhalb bestimmter gesellschaftlicher Gruppen, zu erforschen – und umgekehrt, wie Sprache Gesellschaft, Gesellschaften und Gesellschaftsschichten konstituiert. Sowohl Gesellschaft wie auch

³⁶ Pelz, Linguistik. Einführung (wie Anm. 31), 24.

³⁷ Vgl. neben Vater, Einf. Sprachwissenschaft (wie Anm. 31), 230ff. auch: Holger Kuße, Kulturwissenschaftliche Linguistik. Eine Einführung (Göttingen 2012), 77ff.

Sprache sind innerhalb ihrer selbst nicht homogen, sondern wesentlich differenziert. Gesellschaftliche und sprachliche Differenzierungen bedingen einander. Zu den essentiellen Forschungsgegenständen der Soziolinguistik zählt daher die Sprachlichkeit verschiedener sozialer Gruppen innerhalb einer Sprachgemeinschaft, das heißt: Sie untersucht verschiedene sprachliche Varietäten entlang gesellschaftlicher Binnendifferenzierung. Dialekte und Soziolekte gehören in diesen Bereich. Soziolinguistik begreift „Einzelsprache als Varietätengefüge“³⁸. Von diesem Varietätengefüge kommen in konkreten kommunikativen Situationen aus jeweils unterschiedlichen Gründen jeweils bestimmte Teile zur sprachlichen Anwendung und andere wiederum nicht (*parole*). Soziolinguistik fragt nach den „soziosituativen Bedingungen für die Verwendung bestimmter [sprachlicher] Varietäten oder Merkmalgruppen innerhalb einer bestimmten Kommunikationsgemeinschaft“³⁹ sowie nach den Funktionen dieser Varietäten.

Soziolinguistik fragt auch, was Sprachgebrauch über die Beziehungen der Sprecher untereinander aussagt. Ebenso werden von Soziolinguistinnen und -linguisten Phänomene untersucht wie sprachlicher Wandel, Sprachkontakte, woraus sich Überschneidungen mit der Kontaktlinguistik ergeben, Sprachmischung und Sprachkonflikte. Die Erforschung von geschlechterspezifischer Sprache bildet mittlerweile eine eigene Teildisziplin der Sprachwissenschaft, die Genderlinguistik.

Pointiert könnte man die Grundfragen der Soziolinguistik etwa wie folgt zusammenfassen: Wer spricht gegenwärtig was in welcher Sprache warum mit wem mit welchen sprachlichen Mitteln und unter welchen sozialen Umständen, mit welchen Absichten und mit welchen sozialen Konsequenzen?⁴⁰ Näherhin soll das Phänomen Soziolinguistik hier nicht erläutert werden, allenfalls dann, wenn wir in Folge auf *historische* Soziolinguistik zu sprechen kommen. Bei der Soziolinguistik handelt es sich um ein extrem breites und sehr offenes Forschungsfeld mit zahlreichen Überschneidungen zur Soziologie. Die Literatur

³⁸ Vgl. Klaus *Mattheier*, Historische Soziolinguistik. Ein Forschungsansatz für eine künftige europäische Sprachgeschichte. In: Helga *Bister-Broosen* (Hg.), Beiträge zur historischen Stadtsprachenforschung (Schriften zur diachronen Sprachwissenschaft 8, Wien 1999), 223–234; hier: 226.

³⁹ *Mattheier*, Hist. Soziolinguistik (wie Anm. 38), 227.

⁴⁰ Vgl. Joshua A. *Fishman*, The Sociology of Language. An Interdisciplinary Social Science Approach to Language in Society (Rowley/Mass. 1972), 15.

zur Soziolinguistik gilt als „unüberschaubar“.⁴¹ Klaus Mattheier weist nicht ganz ohne polemischen Unterton darauf hin, dass eine extrem breite Definition von Soziolinguistik, „nach der die gesamte Sprachlichkeit rückgebunden ist an gesellschaftliches Handeln“⁴², dazu führt, dass praktisch die gesamte Linguistik als Soziolinguistik aufgefasst werden müsse.

1.2.3) Historische Soziolinguistik

Die historische Soziolinguistik wendet die Grundfragen der Soziolinguistik auf Sprachgemeinschaften vergangener Tage an. Historische Soziolinguistik beschäftigt sich demnach, auf den Punkt gebracht, mit „*past language states*“⁴³, also mit historischen Sprachzuständen, sprachlich generierten Beziehungen und Sprechakten der Vergangenheit – und zwar unter ausdrücklicher Berücksichtigung der Wechselbeziehung von Sprache mit kultur- und mit sozialgeschichtlichen Gegebenheiten. Vornehmlich untersucht sie die Beschaffenheit historischer Sprachstufen diachron und synchron, deren Veränderungen im Laufe der Zeit (*variation; language change*) sowie die soziokulturellen Hintergründe und Voraussetzungen dafür, dass sprachliche Veränderungen sowie bestimmte sprachliche Zustände eingetreten sind (und: in welcher Weise sie eingetreten sind). Sie fragt beispielsweise: Welcher Sprachgebrauch zeichnet eine bestimmte soziale Gruppe der Vergangenheit aus? Wer spricht warum mit wem welche sprachliche Varietät in welchen Situationen (Sprachgebrauchsgeschichte)? Welche kommunikativen und welche sozialen Funktionen und Implikationen hatte ein bestimmter gruppenspezifischer Sprachgebrauch, welche sozialen Konsequenzen innerhalb, welche Folgen außerhalb einer Gruppe? Wie und warum wurden bestimmte sprachliche Erscheinungen zum Zeitpunkt X von speziellen sozialen Gruppierungen als solche wahrgenommen, warum und wie gebraucht oder auch nicht wahrgenommen, nicht gebraucht, wie beurteilt? Unter welchen Voraussetzungen verbreitete ein sprachliches Phänomen sich (nicht) in einer bestimmten

⁴¹ Schlobinski, Grundfragen Sprachwissenschaft (wie Anm. 33), 79.

⁴² Mattheier, Hist. Soziolinguistik (wie Anm. 38), 223.

⁴³ Vgl. Paul T. Roberge, Language History and Historical Sociolinguistics. Sprachgeschichte und historische Soziolinguistik. In: Ulrich Ammon, Norbert Dittmar, Klaus Mattheier, Peter Trudgill (Hgg.), Sociolinguistics/Soziolinguistik. An International Handbook of the Science of Language and Society/Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science, Bd. 3, 3. Teilband., Berlin 2006), 2306–2315; hier: 2311.

historischen Sprachgemeinschaft (oder auch darüber hinaus)? Unter welchen sozialen Voraussetzungen wurde eine sprachliche Veränderung in der Vergangenheit zur standardsprachlichen Norm? Wie lange dauerte dieser Vorgang der Inkulturation des Standards? Warum schafften es manche Phänomene in die Hochsprache und warum verschwanden andere wieder? Wie hat der Sprachgebrauch einer spezifischen Gruppen auf den einer anderen Gruppe eingewirkt (Sprachkontaktgeschichte)? – Die historische Soziolinguistik fragt nach sozialgeschichtlichen Rahmenbedingungen und Konsequenzen von sprachlichen Phänomenen, von Sprachgebrauch, Sprachkontakt und von Sprachwandel in der Vergangenheit und ist bemüht um eine „alltagsgeschichtliche und kulturhistorische Einbettung von Sprach(gebrauchs)geschichte. [...] Es geht [...] auch um die Deutung soziokultureller Diversität mit den Mitteln der Sprache, das heißt um die Frage, ob und wie die Sprachgebräuche gesellschaftlicher Gruppen der symbolischen Konstruktion sozialer Ordnungen dienen.“⁴⁴

„[In] order to understand linguistic change [or variation], one must see it as a part of a wider process of cultural change.“⁴⁵ Die historische Soziolinguistik weist (im Gegensatz zur rein historischen Linguistik) nachdrücklich darauf hin, dass sprachliche Veränderung sich nie unabhängig von gesellschaftlichen Gegebenheiten und Veränderungen abspielt (und vice versa), sondern dass das eine auf das Engste mit dem anderen verknüpft ist. Sprachlicher und sozialer Wandel gehen miteinander Hand in Hand. Die historische Soziolinguistik betont, dass sprachliche Veränderung vom Sprecher ausgeht. Sie nimmt also das Subjekt dezidiert als historisch-linguistischen Akteur wahr. Dabei wendet sie Theorien und (vor allem quantitative) Methoden der Soziolinguistik an. Insgesamt muss sie auf schriftliche Überlieferung zurückgreifen. „The objective is the extrapolation of patterns of [linguistic] variation from written source material and/or the documentation of linguistic changes across a stipulated temporal range, and their correlation with other linguistic and extralinguistic phenomena.“⁴⁶

⁴⁴ Angelika Linke, „Wer sprach warum wie zu einer bestimmten Zeit?“ Überlegungen zur Gretchenfrage der Historischen Soziolinguistik am Beispiel des Kommunikationsmusters ‚Scherzen‘ im 18. Jahrhundert. In: Ulrich Ammon, Jeroen Darquennes, Leigh Oakes, Sue Wright (Hgg.), *Sociolinguistica. Internationales Jahrbuch für europäische Soziolinguistik*, Bd. 13, Heft 1 (1999), 179–208; hier: 180f.

⁴⁵ Roberge, *Language History and Historical Sociolinguistics* (wie Anm. 43), 2307.

⁴⁶ Roberge, *Language History and Historical Sociolinguistics* (wie Anm. 43), 2311.

Ursprünge und Postulat eines sprachwissenschaftlichen Programms, das historische und sozialgeschichtliche Komponenten miteinbezieht, liegen in den 1960er-Jahren, doch konnte Klaus Mattheier für den deutschsprachigen Raum noch in den späten 1980er-Jahre konstatieren, dass „die allgemeine Sprachwissenschaft [...] sich auch heute in der Regel noch nicht den Fragestellungen des Verhältnisses zwischen Sprache, Geschichte und Gesellschaft [stellt]. Eine Historische Soziolinguistik gibt es noch nicht, und wenn es sie gäbe, so wäre noch nicht sicher, ob sie tatsächlich innerhalb der Sprachwissenschaft anzusiedeln wäre oder [...] in der Soziologie oder in der Geschichtswissenschaft.“⁴⁷

In weiterer Folge entfaltete sich die historische Soziolinguistik langsam ab den 1990er-Jahren, und das vor allem im deutsch-, niederländisch- und englischsprachigen Raum. Dennoch konstatieren Roland Willemyns und Wim Vandenbussche noch im Jahr 2006, die historische Soziolinguistik sei zu Beginn der 2000er-Jahre eine Disziplin gewesen „still quite uncertain on how to proceed. [...] Until the end of the nineties the amount of books and articles referring to historical sociolinguistics in their title is rather low [...]“.⁴⁸ Seitdem hat die historische Soziolinguistik einen regelrechten Boom erlebt und sich auf viele Sprachen erweitert. Gleichwohl beklagten Willemyns und Vandenbussche 2006 die Beschränkung der Forschungsbestrebungen auf die jeweils einzelne Nationalsprache und monieren die Vernachlässigung einer gemeinsamen Perspektive auf der Ebene einer gesamteuropäischen historischen Soziolinguistik. Latein als Gegenstand linguistisch forschender Historiker*innen scheint, abgesehen von den Publikationen Peter Burkes, Roy Porters und Françoise Waquets (siehe Literaturverzeichnis), eine eher untergeordnete Rolle zu spielen.

⁴⁷ Klaus *Mattheier*, Nationalsprachenentwicklung, Sprachenstandardisierung und Historische Soziolinguistik. In: Ulrich *Ammon*, Klaus *Mattheier*, Peter *Nelde* (Hgg.), *Sociolinguistica. Internationales Jahrbuch für europäische Soziolinguistik*, Bd. 2, Heft 1 (1988), 1–9; hier: 1.

⁴⁸ Wim *Vandenbussche*, Roland *Willemyns*: Historical Sociolinguistics. Coming of Age? In: Ulrich *Ammon*, Klaus *Mattheier*, Peter *Nelde* (Hgg.), *Sociolinguistica. Internationales Jahrbuch für Europäische Soziolinguistik*, Bd. 20 (2006), 146–165; hier: 146 & 157.

1.2.4) Computerlinguistik und Texttechnologie

Einer gängigen, weil eingängigen Definition zufolge, die als möglichst allgemein und als umfassend verstanden werden will, ist Computerlinguistik jene Disziplin „im Überschneidungsbereich von Informatik und Linguistik“⁴⁹, die sich mit der maschinellen Verarbeitung, mit der maschinellen Analyse sowie mit der maschinellen Beschreibung natürlicher Sprachen (im Gegensatz zu Programmiersprachen) beschäftigt. Gegenstand der Untersuchung ist dabei sowohl geschriebene wie auch gesprochene Sprache. Textcorpora dienen dabei häufig als empirische Grundlage. Grenzen zwischen Corpus- und Computerlinguistik erweisen sich somit vielfach als fließend. Beide Disziplinen forcieren einen *quantitativen* Zugang zum Phänomen Sprache. Dies bringt einerseits die Einbeziehung statistischer Methoden mit sich, andererseits wird die computer-gestützte Verarbeitung selbst großer Datenmengen wesentlich begünstigt.⁵⁰

„Computational Linguistics aims at designing, implementing, and applying computational models for natural languages. This includes developing computer systems that automatically process languages (from a morphological, syntactic, semantic, or pragmatic point of view), as well as building, enriching, and using corpus data (Corpus Linguistics), and much more.“⁵¹

Innerhalb der Computerlinguistik, einer verhältnismäßig jungen Disziplin, herrscht bzw. herrschte mindestens bis zu Beginn der 2010er-Jahre eine rege Diskussion darüber, wie das Fach eigentlich zu definieren und zu verstehen sei, wofür hier auf einschlägige einführende Literatur verwiesen werden soll, ebenso in Bezug auf die wichtigsten Nachbardisziplinen der Computerlinguistik.⁵²

Ihrem Ursprung und ihrer Geschichte nach war und ist man innerhalb der Computerlinguistik unter anderem besonders daran interessiert, geschriebene und/oder gesprochene Texte mit maschineller Unterstützung automatisch von einer Sprache in eine andere zu übertragen (maschinelle Übersetzungen), also maschinell zu analysieren und zu transformieren.⁵³ Unabdingbare Voraussetzung

⁴⁹ Carstensen, Ebert et al., Computerlinguistik und Sprachtechnologie (wie Anm. 21), 1.

⁵⁰ Vgl. Barbara McGillivray, *Methods in Latin Computational Linguistics* (Bosten/Leiden 2014), 5.

⁵¹ McGillivray, *Methods in Latin Computational Linguistics* (wie Anm. 50), 1.

⁵² Vgl. Carstensen, Ebert et al., Computerlinguistik und Sprachtechnologie (wie Anm. 21), 2–6; vgl. ferner: Lobin, Computerlinguistik (wie Anm. 28), 11–13 & 20.

⁵³ Vgl. Lobin, Computerlinguistik (wie Anm. 28), 14; ferner: Lothar Lemnitzer, Heike

dafür war und ist nach wie vor, dass das natürlichsprachliche Rohdatenmaterial entsprechend abstrahiert und formalisiert, also *modelliert*, das heißt: in eine maschinell lesbare Form gebracht wird, die eine Untersuchung und eine Weiterverarbeitung mit dem Computer überhaupt erst ermöglicht (Modellierung).⁵⁴ Stets folgt die Modellierung dabei bestimmten formalen Beschreibungs- bzw. Aufbauschemata von Sprache (Grammatiken; näherhin unter anderem: Dependenz- und/oder Konstituentengrammatik⁵⁵); und sie folgt bestimmten Algorithmen, also detailliert festgelegten Prozeduren zur Lösung eines Problems, wobei Algorithmen „aus diskreten Schritten bestehen und endlich beschreibbar sein“⁵⁶ sollten.

Innerhalb der Computerlinguistik stellt die Texttechnologie einen relativ jungen Teilbereich dar, der „sich mit der linguistisch motivierten Informationsanreicherung und Verarbeitung digital verfügbarer Texte mittels standardisierter Auszeichnungssprachen beschäftigt.“⁵⁷ Hier bzw. im Bereich des *natural language processing* (NLP) ist digitale Annotation streng genommen zu verorten. Zum Begriff der Auszeichnungssprache werden wir etwas später ein prominentes Beispiel besprechen, nämlich XML.

Zinsmeister, Korpuslinguistik. Eine Einführung (Tübingen ³2015), 71–80; zu weiteren praktischen Anwendungen der Computerlinguistik vgl. Carstensen, Ebert et al., Computerlinguistik und Sprachtechnologie (wie Anm. 21), 553–658.

⁵⁴ Vgl. Thaller, Digital Humanities als Wissenschaft. In: Jannidis, Kohle, Rehbein (Hgg.), Digital Humanities (wie Anm. 23), 13–18; hier: 16.

⁵⁵ Vgl. Lobin, Computerlinguistik (wie Anm. 28), 25–33.

⁵⁶ Ralf Klabunde, Automatentheorie und Formale Sprachen. In: Carstensen, Ebert et al., Computerlinguistik und Sprachtechnologie (wie Anm. 21), 66–93; hier: 67.

⁵⁷ Georg Rehm, Texttechnologische Grundlagen. In: Carstensen, Ebert et al., Computerlinguistik und Sprachtechnologie (wie Anm. 21), 159–168; hier: 159.

1.2.5) Corpuslinguistik und sprachliche Corpora

Corpuslinguistik stellt innerhalb der Sprachwissenschaft eine empirische Teildisziplin bzw. Methode dar,⁵⁸ die darauf basiert und abzielt, natürliche Sprachen bzw. deren Gebrauch an Hand von solchen authentischen und natürlichen Sprachdaten zu beschreiben und zu untersuchen, „die in Korpora zusammengefasst sind.“⁵⁹ Der Begriff Corpus bezeichnet in diesem Zusammenhang eine nach bestimmten Kriterien definierte Teilmenge aus einer übergeordneten Gesamtmenge von Daten, die in natürlicher Sprache kodiert sind.⁶⁰ Diese Sprachdaten respektive Corpora können ein- oder mehrsprachig sein, entweder in schriftlicher Form vorliegen, als gesprochene Texte in Form von Audio-Aufzeichnungen oder auch multimedial.⁶¹ Wenn in der Folge demnach von Sprachdaten bzw. von Corpora die Rede ist, so sind damit ausschließlich natürliche Sprachdaten schriftlichen Niederschlags gemeint. In der Praxis handelt

⁵⁸ Zur Geschichte des Faches bzw. zur Diskussion, ob, inwiefern und ab wann die Corpuslinguistik sich von einer Methode zur eigenständigen Teildisziplin entwickelte, vgl. Susanne Lenz, *Korpuslinguistik* (Tübingen 2000), 6f.; zur Corpuslinguistik als eigenständiger Methodologie und zu den Ansätzen corpusbasiert (*corpus-based*) im Gegensatz zu corpusgeleitet (*corpus-driven*) vgl. e.g. Holger Keibel, Marc Kupietz, Rainer Perkuhn, *Korpuslinguistik* (Paderborn 2012), 18ff.; wenngleich wir Überlegungen in diese Richtung im Rahmen der vorliegenden Arbeit nicht voll ausformulieren können, so neigen wir doch dem corpusgeleiteten Ansatz zu (*corpus-driven*), dem gemäß Daten eines Corpus „nicht dazu [dienen], erst im Nachhinein Thesen und Theorien zu bestätigen oder zu widerlegen, [sondern vielmehr] den Ausgangspunkt [bilden], von dem aus Thesen abgeleitet und Theorien aufgestellt werden.“ – Keibel, Kupietz, Perkuhn, *Korpuslinguistik*, 20. „Das Ziel der korpusgeleiteten Methode besteht darin, eine Theorie zu erstellen, die in Einklang mit den Korpusdaten steht.“ – Ilka Mindt, *Methoden der Korpuslinguistik. Der korpusbasierte und der korpusgeleitete Ansatz*. In: Iva Kratochvílová, Norbert Richard Wolf (Hgg.), *Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive* (Heidelberg 2010), 53–65; hier: 54. Dem gegenüber ist es „[d]as Ziel der korpusbasierten Methode [...], Korpusdaten zur nachträglichen Erklärung, Veranschaulichung und/oder Überprüfung einer [a priori bestehenden] Theorie zu verwenden.“ – Mindt, *Methoden* (w.o.), ebd.; einen erweiterten und eigenständigen Begriff von corpusgeleitet vertreten Gard B. Jensen, Barbara McGillivray, *Quantitative Historical Linguistics* (Oxford Studies in Diachronic and Historical Linguistics 26, Oxford 2017), 58 ff.

⁵⁹ Lemnitzer, Zinsmeister, *Korpuslinguistik* (wie Anm. 53), 14; einen Gegensatz zu natürlichen Sprachen bilden beispielsweise Programmiersprachen.

⁶⁰ Im Sinne eines erweiterten geschichtswissenschaftlichen Quellenbegriffes, der sich nicht nur auf Schrift- bzw. Sprachdenkmäler beschränkt, ist an dieser Stelle kurz darauf hinzuweisen, dass Daten bzw. Quellen nicht zwangsläufig sprachlicher Natur sein müssen, sondern etwa als Bilder auch optische Qualität aufweisen können, dass es entsprechend auch Bildcorpora geben kann oder Sammlungen von dinglichen Quellen, was jedoch vom Bereich der Corpuslinguistik wegführt und in die Kompetenz der historischen Nachbardisziplinen fällt, wie etwa der Archäologie oder der Kunstgeschichte.

⁶¹ Eine eingehendere Klassifikation verschieden gearteter Corpora bietet Scherer, *Korpuslinguistik* (wie Anm. 30), 16–31. Auf Corpora in Form von Audio-Aufzeichnungen sowie auf multimediale Corpora wird im Kontext vorliegender Arbeit nicht eingegangen.

es sich dabei meist um Sprachdaten lebender Sprachen. Textcorpora historischer Sprachstufen liegen im Vergleich zu modernsprachlichen in weit geringerer Anzahl vor. Die Zielsetzung für die Erstellung von Corpora kann es entweder sein, den Sprachgebrauch einer (National-)Sprache oder einer historischen Sprachstufe an sich in Grundzügen abzubilden (Referenzcorpora), um auf dieser Basis allgemeine Aussagen über eine Sprache zu treffen, wonach ein Querschnitt aus verschiedenen Textsorten einbezogen werden muss, oder es handelt sich um Spezialcorpora unterschiedlicher Schwerpunkte (z.B. Corpora mit Sprachdaten von Sprecherinnen und Sprechern einer bestimmten Altersgruppe, einer sozialen Schicht, eines bestimmten Geschlechtes et c.).

Charakteristisch nicht nur für ein sprachwissenschaftliches Corpus, sondern für Textcorpora generell, ist es nun, dass sie, wie bereits angedeutet, nach außen hin nach bestimmten Kriterien gegenüber größeren Gesamtmengen von Sprachdaten abgegrenzt wurden. Hinzu kommt noch, dass sie nach innen hin nach bestimmten Kriterien strukturiert und geordnet sind. Für (linguistische) Corpora gilt demnach: „Ein Korpus ist eine Sammlung von [gesprochenen oder geschriebenen] Texten oder Textteilen, die bewusst nach [vorab] bestimmten sprachwissenschaftlichen [oder anderen] Kriterien ausgewählt und geordnet [wurden]“⁶². Die erwähnten Kriterien können sich dabei prinzipiell nach der Art der Fragestellung richten, die man mittels Erstellung und Untersuchung eines Corpus bearbeiten möchte, zumal „ein Korpus immer im Hinblick auf einen bestimmten Verwendungszweck erstellt wird.“⁶³ Von der Computerlinguistik werden Corpora häufig als empirische Datengrundlage genutzt, an Hand derer Programme oder Software zur automatischen Sprachverarbeitung trainiert werden.⁶⁴

Der Begriff des (Text-)Corpus im Sinne einer nach bestimmten wissenschaftlichen Kriterien erstellten (Text-)Sammlung kann generell jedoch nicht auf die Linguistik alleine beschränkt werden. Daher wird man im Rahmen der entsprechenden Disziplinen auch von historisch-quellenkundlichen oder von bestimmten literaturwissenschaftlichen Corpora sprechen dürfen. Die Notwendigkeit, ein Textcorpus zu erstellen, ergibt sich nicht nur innerhalb der Corpuslinguistik alleine, sondern

⁶² Scherer, *Korpuslinguistik* (wie Anm. 30), 3.

⁶³ Scherer, *Korpuslinguistik* (wie Anm. 30), 16.

⁶⁴ Vgl. Heike Zinsmeister, *Korpora*. In: Carstensen, Ebert et al., *Computerlinguistik und Sprachtechnologie* (wie Anm. 21), 482–491; hier: 482.

überall dort, wo es etwa aus arbeitsökonomischen Gründen nicht mehr sinnvoll und nicht mehr möglich erscheint, die gesamte Masse an geschichts-, literatur- oder sprachwissenschaftlichen Quellen, die zur Verfügung stehen, zu überblicken und zu bearbeiten, so etwa bei der Sammlung oder Edition historischer Periodika. Dass Ursprung und Verwendungszweck eines Textcorpus nicht zwangsläufig und nicht ausschließlich innerhalb der Sprachwissenschaft liegen müssen, versteht sich von selbst.⁶⁵ Selbstverständlich können auch historische Fragestellungen den Ausgangspunkt für die Erstellung von Textcorpora bilden.

Hierbei ist klar, dass die inneren und die äußeren Eigenschaften eines Corpus wesentlich von folgenden Faktoren abhängen: der Forschungsdisziplin, der zu Grunde liegenden Fragestellung und den definierten Erstellungskriterien. Hinsichtlich der daraus resultierenden Eigenschaften des Textcorpus generell ist es in jeder wissenschaftlichen Disziplin von zentraler Bedeutung, dass das erstellte Corpus einen repräsentativen Ausschnitt jenes Gegenstandes darstellt, der untersucht werden soll. Sei dieser Untersuchungsgegenstand nun Sprache selbst, eine bestimmte Quellen- oder Literaturgattung oder etwa das Kanzleiwesen eines spätmittelalterlichen Kaisers des Hl. Römischen Reiches:

So man, wie gesagt, nicht die gesamte Masse an historischen, literatur- oder sprachwissenschaftlichen Quellen zu bearbeiten im Stande ist, ergibt sich zwangsläufig die Notwendigkeit, repräsentative Quellencorpora zu erstellen, die den eigenen Fragestellungen entgegenkommen, je nach akademischer Disziplin nach unterschiedlichen Kriterien.

Unabhängig davon, in welchem Forschungsbereich man nun konkret tätig ist, hat Repräsentativität als *die* zentrale Eigenschaft des Quellencorpus bei seiner Erstellung oberste Priorität. Schließlich soll ein Quellencorpus trotz oder gerade wegen seiner quantitativen Beschränktheit einen größeren Sachverhalt abbilden, sei dieser Sachverhalt nun von geschichts-, sprach- oder von literaturwissenschaftlichem Interesse. Allerdings: Nicht ganz zu Unrecht wird innerhalb der Corpuslinguistik fallweise die Ansicht vertreten, Repräsentativität zu erreichen sei angesichts der Heterogenität des sprachwissenschaftlich relevanten Materials eine nicht zu realisierende Idealvorstellung, da es „schlichtweg nicht möglich [sei], den

⁶⁵ Ob es innerhalb der Geschichtswissenschaft sinnvoll ist, Textcorpora ausschließlich nach sprachwissenschaftlichen Kriterien zu definieren, wäre demnach zu diskutieren.

Sprachgebrauch als Ganzes zu erfassen, und die Handhabbarkeit der Korpora auch deren Größe beschränken kann [...].“⁶⁶ Insofern erscheint es zweckmäßig, Korpora als *möglichst* adäquate und ausgewogene Modelle von Sprachgebrauch anzusehen und anzustreben.⁶⁷

In unmittelbarem Zusammenhang mit der Repräsentativität eines Corpus stehen im Zuge seiner Erstellung Parameter wie Größe und Inhalt. Für diese Parameter existieren keine allgemein verbindlichen Maßstäbe. Vielmehr ergeben Größe und Inhalt sich aus Forschungsgegenstand und Forschungsfrage, der bzw. die einem Corpus zu Grunde liegt. Die Frage, ab welcher Quantität ein Corpus als repräsentativ für einen bestimmten Gegenstand angesehen werden kann, kann nicht pauschal beantwortet werden. „Ein Korpus ist nicht von vornherein groß.“⁶⁸ Große wie kleine Korpora bieten beiderseits Vor- und Nachteile. Umfang und Inhalt stellen bei der Kompilation von Korpora deshalb grundsätzlich variable Größen dar. Entsprechend dem Forschungszweck und dem Forschungsgegenstand, der einem Corpus zu Grunde liegt, weist der Inhalt der ausgewählten Texte unterschiedlich hohe Ausmaße an Homo- bzw. Heterogenität auf. Prinzipiell hat die Auswahl der Texte jedoch nach vorab festgelegten, transparenten und intersubjektiv überprüfbaren Kriterien zu erfolgen, die ihrerseits eine sinnvolle Struktur aufweisen müssen.

Einmal aufgebaut, können Größe, Inhalt und Struktur eines Corpus unverändert bleiben, müssen dies aber nicht zwangsläufig. Das Kriterium der Beständigkeit stellt keine *condicio sine qua non* für ein Corpus dar. Korpora müssen nicht per se statisch bleiben. In manchen Fällen mag es sich als sinnvoll erweisen, einmal konstituierte Korpora prinzipiell offen zu halten oder nach einem ersten Abschluss auch nachträglich um zusätzliche Texte zu erweitern (z.B. Monitorcorpora). Dies trifft unter Beibehaltung vornehmlich von Inhalt und Struktur des Corpus vor allem in quantitativer Hinsicht zu, wenn sich herausstellt, dass ein Text kraft seiner Eigenschaften die Repräsentativität eines Corpus in Hinblick auf einen bestimmten Sachverhalt, in Hinblick auf den dem Corpus zu Grunde liegenden

⁶⁶ Noah Bubenhofer, Marek Konopka, Roman Schneider, Präliminarien einer Korpusgrammatik (Tübingen 2014), 46.

⁶⁷ Vgl. Bubenhofer, Konopka, Schneider, Präliminarien (wie Anm. 66), 46.

⁶⁸ Norbert Richard Wolf, Korpora in der Korpuslinguistik. In: Kratochvílová, Wolf, Kompendium (wie Anm. 58), 17–25; hier: 23.

Forschungsgegenstand wesentlich erweitert, jedoch noch nicht Teil des Corpus ist. Vom Faktor der Beständigkeit ist der Faktor der Abgeschlossenheit prinzipiell zu unterscheiden. Ich möchte den Begriff Beständigkeit im vorliegenden Zusammenhang ausschließlich quantitativ verstanden wissen, den Begriff der Abgeschlossenheit hingegen rein qualitativ. Denn insofern, als jedes Corpus eine nach bestimmten Kriterien definierte Teilmenge aus einer übergeordneten Gesamtmenge von Daten darstellt, ist es gegenüber dieser Gesamtmenge in qualitativer Hinsicht zwar abgeschlossen, und diese qualitative Abgeschlossenheit – hinsichtlich der fundamentalen Kriterien, hinsichtlich Inhalt, Aufbau und Struktur des Corpus – sollte aus arbeitsökonomischen Gründen auch tunlichst beibehalten werden, aber: Qualitative Abgeschlossenheit bringt eine quantitative Beständigkeit gerade nicht zwangsläufig mit sich. Stellt sich nämlich nach einem ersten Abschluss des Corpus heraus, dass ein darin noch nicht aufgenommener Text seinen Kriterien prinzipiell entspricht, so kann grundsätzlich in Erwägung gezogen werden, ihn im Nachhinein in das Corpus aufzunehmen. Wenn Scherer also davon spricht, ein Corpus erfülle notwendigerweise „das Kriterium Beständigkeit“⁶⁹, so gilt dies vor allem in qualitativer Hinsicht, im Sinne der hier skizzierten Abgeschlossenheit qua vordefinierter Kriterien, nicht jedoch prinzipiell in Hinblick auf das Textquantum, das im Corpus enthalten ist.

Bezüglich weiterer fundamentaler Eigenschaften von (linguistischen) Textcorpora scheint sich innerhalb der Sprachwissenschaft im Laufe der letzten Jahrzehnte eine rege Diskussion darüber entwickelt zu haben, inwiefern es für Corpora konstitutiv sei, dass sie maschinenlesbar sind, in digitaler Form vorliegen, ebenso erschlossen und angereichert sind durch computergestützte linguistische Annotation sowie durch die sich (nicht nur, aber vor allem) daraus ergebenden (digitalen) Metadaten. Die Details dieser Diskussion können im Rahmen der vorliegenden Arbeit nicht nachvollzogen werden. Es scheint jedoch so, als würde der Corpus-Begriff von etlichen Linguistinnen und Linguisten nur noch dann akzeptiert, wenn die (Sprach-)Daten eines Corpus in digitaler bzw. digitalisierter Form vorliegen, maschinenlesbar und auf Rechnern gespeichert sind sowie unter

⁶⁹ Scherer, Korpuslinguistik (wie Anm. 30), 6.

Zuhilfenahme texttechnologischer Methoden linguistisch annotiert und durch (digitale) Metadaten erschlossen wurden.⁷⁰

Vertritt man im traditionellen Sinne einen weiteren Corpus-Begriff, wie wir ihn soeben skizziert haben, so fallen darunter auch „eine Vielzahl von Textsammlungen, die nicht computerlesbar sind.“⁷¹ So sehr ich im Anschluss an Scherer diesem traditionellen Verständnis von Corpus als Oberbegriff für bewusst nach wissenschaftlichen Kriterien erstellte Quellensammlungen zuneige, so problematisch ist diese Definition doch für die Praxis. Vertritt man nämlich einen traditionellen Corpus-Begriff, so ergibt sich innerhalb seiner selbst die Notwendigkeit einer sachlichen und terminologischen Binnendifferenzierung. Nicht computerlesbare Corpora müssen gegenüber computerlesbaren abgegrenzt werden. Erstere werden zu diesem Zweck „häufig als Belegsammlungen, Textarchive oder ähnliches bezeichnet.“⁷²

Diese Terminologie führt, zumal innerhalb der Editionswissenschaften, zu Problemen. Denn: Es stellt sich die Frage nach der exakten Definition und der exakten Verwendung dieser Termini. Die Ausdrücke Textarchiv oder Belegsammlung sind, soweit ich derzeit sehe, weder in der Sprachwissenschaft noch im Editionswesen klar umrissen. Es stellt sich innerhalb des Editionswesens zudem die Frage nach der Anwendung des Begriffs Corpus auf verschiedene Bearbeitungsstufen eines Textes, der digital oder traditionell-analog ediert werden soll. Ein Beispiel: Innerhalb der Gesamtmenge neulateinischer Texte aus dem 18. Jahrhundert könnte die Gattung der Gelehrtenbriefe als Corpus angesehen werden. Sie kann zusätzlich eingegrenzt werden, etwa nach regionalen oder nach konfessionellen oder nach inhaltlichen Kriterien oder nach Sender und Empfänger. Innerhalb der Gattung neulateinischer Gelehrtenbriefe aus der genannten Zeit kann demnach die Korrespondenz der Gebrüder Pez als Corpus angesehen werden und innerhalb dieser Korrespondenz wiederum eine gewisse Teilmenge daraus als eigenes (Teil-)Corpus. Nun liegt aber diese Korrespondenz gegenwärtig (Stand Frühjahr 2019) in verschiedenen Bearbeitungsstufen vor, einerseits in Form der originalen und ursprünglichen Handschriften in

⁷⁰ Vgl. Felix Sasaki, Andreas Witt, Linguistische Korpora. In: Lothar Lemnitzer, Henning Lobin (Hgg.), *Texttechnologie. Perspektiven und Anwendungen* (Tübingen 2004), 195–216.

⁷¹ Scherer, *Korpuslinguistik* (wie Anm. 30), 17.

⁷² Scherer, *Korpuslinguistik* (wie Anm. 30), 18.

verschiedenen Archiven und Aufbewahrungsorten, andererseits als gedruckte (Teil-)Edition und teils in digitalen Vorstufen zu einer zukünftigen digitalen Edition. Da die gedruckte Edition gegenwärtig jedoch noch nicht den gesamten Bestand an überlieferten Handschriften umfasst, stellt sich ohne Eingrenzung an Hand strikter Kriterien die Frage: Was daraus ist das „eigentliche“ Corpus? Eine exakte terminologische und sachliche Definition respektive Differenzierung an Hand der vorgebrachten Begriffe Corpus, Belegsammlung und Textarchiv alleine erscheint nicht sinnvoll. Die Kriterien der Unterscheidung müssen andere sein. Die Definition darf nicht auf einer rein terminologischen Ebene stattfinden, sondern sie muss wissenschaftlichen und exakten Maßstäben genügen.

Der Oberbegriff des Corpus für eine bewusst nach wissenschaftlichen Kriterien erstellte Quellensammlung, gleich ob computerlesbar oder nicht, kann jedoch auch innerhalb des Editionswesens beibehalten werden, wenn man ausdrücklich darauf hinweist, dass eine Quelle kein Corpus in diesem Sinne darstellt, eine Edition nur mit Einschränkungen. Eine Edition, so sie sich nicht ausdrücklich als Teiledition versteht, strebt nach der vollständigen Wiedergabe von Quellentexten. Sie soll „in erster Linie einen zuverlässigen Text zur Verfügung stellen [...], der die Grundlage jedweder historischen und interpretatorischen Betrachtung bildet.“⁷³ Insofern eine vollständige Edition tendenziell aber *keinen Ausschnitt* aus einer Gesamtmenge an Quellentexten darstellen will, sondern – traditionell gesehen – einen Anspruch auf „Vollständigkeit“ zu stellen scheint, ist sie kein Corpus im strengen Sinne. Sie verfolgt auch keine linguistischen Ziele – jedenfalls nicht primär. Und umgekehrt stellt die unbearbeitete Quelle ebenso wenig ein Corpus dar, solange sie nicht den Prozess der wissenschaftlichen Bearbeitung durchlaufen hat. Der Zufall der Überlieferung, der dazu führt, dass erhaltene Quellen eben erhalten sind, kann nicht als Kriterium für die Definition eines Corpus dienen. Folglich ist es im Editionswesen zwingend erforderlich, die Begriffe Quelle, Edition und Corpus terminologisch deutlich voneinander abzugrenzen.

Indes wird der Begriff der digitalen Edition innerhalb der Computerlinguistik allzu leichtfertig verwendet, zumal aus der Sicht der Editionswissenschaft und aus

⁷³ Bodo *Plachta*, Editionswissenschaft. Eine Einführung in Methode und Praxis der Edition neuerer Texte (Stuttgart/Ditzingen ³2013), 12.

der Sicht der Digital Humanities. Weder bei der Perseus Digital Library (PDL) noch bei der Library of Latin Texts (LLT) noch bei der Bibliotheca Teubneriana Latina (BTL) oder bei der Online-Version des Corpus Grammaticorum Latinorum (CGL) handelt es sich um genuin „digital editions“⁷⁴. Vielmehr wurden im Rahmen dieser Unternehmen solche Editionen *digitalisiert*, die ursprünglich gedruckt vorlagen. Die PDL, die LLT, die BTL und das CGL stellen somit Sammlungen *digitalisierter* Editionen dar, die bei der Digitalisierung freilich mancher Eigenschaften beraubt wurden, über die sie in gedrucktem Zustand noch verfügten. Insofern und im Vergleich mit sprachlich annotierten Corpora mag es auf den ersten Blick sogar berechtigt erscheinen, diese Textsammlungen als „raw text corpora“⁷⁵ zu bezeichnen, zumal annotierte Corpora viel besser mittels digitaler Suchabfrage nach linguistischen Phänomenen durchsucht werden können. Dass aber bei einer zweifelhaften Verwendung des Begriffes der digitalen Edition dieser offenbar mit „raw text corpora“ gleichgesetzt wird, erscheint sowohl aus Sicht der Editionswissenschaft wie auch aus Sicht der Corpuslinguistik und der Digital Humanities korrekturbedürftig.

1.2.6) Digital-linguistische Annotation und XML

Der Begriff (digitale) Annotation⁷⁶ steht für die (computergestützte) Anreicherung eines Rohtextes (Primärdaten), der nicht zwingend, aber meist digitalisiert oder bereits genuin digital vorliegt, mit zusätzlichen Informationen. Diese Informationen dienen seiner Erschließung und seinem Verständnis.⁷⁷ Dem Grundgedanken nach ist das Konzept der Annotation sehr alt. Im Prinzip existiert es, seit Texte schriftlich tradiert, interpretiert und kommentiert werden. Man geht nicht fehl darin, zu sagen, Annotieren zähle „zu den ältesten und grundlegendsten

⁷⁴ McGillivray, *Methods in Latin Computational Linguistics* (wie Anm. 50), 5.

⁷⁵ McGillivray, *Methods in Latin Computational Linguistics* (wie Anm. 50), 7.

⁷⁶ Vgl. Scherer, *Korpuslinguistik* (wie Anm. 30), 21f. & 58f.

⁷⁷ Wenn wir in Folge die wichtigsten Unterarten von Annotation kurz darstellen, so verzichtet diese Darstellung bewusst auf den Anspruch, alle Unterarten vollständig zu repräsentieren. Die Auswahl der dargestellten Annotationsarten erfolgt vielmehr aus Platzgründen und nach dem Kriterium, sich auf das Wesentlichste zu beschränken und dabei möglichst übersichtlich und allgemein verständlich zu bleiben. Ferner wird ausdrücklich darauf hingewiesen, dass in einer Frühphase der vorliegenden Studie die Arbeit direkt in TEI-XML noch nicht in Betracht gezogen wurde, sondern der Fokus ausschließlich auf Annotationswerkzeugen mit graphischer Benutzeroberfläche (graphical user interface – GUI) und auf linguistischer Annotation lag.

[kulturellen] Praktiken“⁷⁸ überhaupt: Einem Text werden zusätzliche Daten und Informationen (Paratexte) beigegeben, die ihn erklären und/oder ergänzen. Man denke in diesem Zusammenhang an Randnotizen, Interlinearglossen und -übersetzungen, die sich in hochmittelalterlichen Handschriften, etwa zu Werken antiker Autoren oder zu sogenannten Heiligen Schriften, finden. Das sind Annotationen. Überspitzt formuliert könnte man die gesamte abendländische, ja weltweite schriftliche Beschäftigung mit und die Deutung von Literatur, von Text, von Informationen an sich (!) als fortwährenden Annotationsprozess bezeichnen. Annotieren dient der Strukturierung, der Anreicherung, der Aneignung, der Erklärung, der Weitergabe und der erklärenden Explizierung von vorhandenen Informationen bzw. von Daten⁷⁹ – Annotieren dient also dazu, den kognitiven Prozess der Informationsverarbeitung und der Wissensaneignung zu erleichtern und zu erweitern.

Im digitalen Bereich können Annotationen auf unterschiedlichsten Ebenen des Textes vorgenommen werden. Man erzeugt sie in der Regel mit digitalen Markierungen. Letzteres gewährleistet, dass der annotierte Text maschinell les-, durchsuch- und analysierbar wird bzw. bleibt. Die hinzugefügten Daten können dabei über den Text selbst hinausgehen und *Metadaten* – beschreibende Daten bzw. Informationen über Daten – enthalten, also etwa: Angaben zum Datenumfang, zur Textgattung, zum Zeitpunkt und zum Ort der Abfassung sowie der Veröffentlichung des Textes, sofern dieser solche Daten nicht selbst enthält, zur Person des Verfassers (z.B. Herkunft, Lebenszeit, Alter, Geschlecht, Bildungsgrad) et c.⁸⁰; oder es handelt sich, um nur *ein* Beispiel zu nennen, um

⁷⁸ Andrea Rapp, Manuelle und automatische Annotation. In: Jannidis, Kohle, Rehbein, Digital Humanities (wie Anm. 22), 253–267; hier: 254.

⁷⁹ Vgl. Rapp, Annotation (wie Anm. 78), 254.

⁸⁰ Keibel, Kupietz, Perkuhn, Korpuslinguistik (wie Anm. 58), 57 f. verwenden den Begriff Metadaten zunächst für Informationen, mit denen ein Corpus als ganzes angereichert wird und unterscheiden Metadaten von Annotationen. Mit dem Begriff Annotationen bezeichnen sie – reichlich unpräzise – „Elemente unterhalb [?] der Textebene“ (ebd., 58) im Sinne gekennzeichnete, linguistisch relevanter Größen. Zu ersteren, den Metadaten, zählen sie zu Recht Angaben über „Zweck, Design [und] Zusammensetzung [des Corpus sowie Angaben über damit] abgedeckte Sprachen bzw. Sprachauschnitte [?] und Nutzungsrechte“ (ebd., 57). Auch gehören Angaben über den Zeitpunkt der Anlage eines Corpus hierher sowie Angaben über seinen Umfang, über seinen Compiler bzw. seine Compiler und schließlich über den Annotationsprozess selbst: Wann wurde er vorgenommen? Welcher Zweck wurde damit verfolgt? Welcher Art ist die Annotation? Nach Erstellung eines Corpus erleichtern es dergleichen Angaben der späteren Forschung, dessen Typ und Inhalt schnell und einfach einzuordnen. Ein äußerst breiten Begriff von Metadaten vertreten Jensen, McGillivray, Quant. Hist. Ling. (wie Anm. 58), 99, wonach im Grunde alles (!), was man über einen Text sagen

Informationen zur grammatikalischen und/oder zur strukturellen Beschaffenheit eines Textes, um analysierende Informationen also, die in einem Text *implizit schon vorhanden sind*, die aber durch die (digitale) Annotierung erst explizit gemacht werden.⁸¹ Annotieren stellt demnach einen (heute zumeist digital unterstützten) wissenschaftlichen Arbeitsprozess dar, an dessen Ende verschieden geartete Annotationen und/oder (Meta-)Daten stehen, die einen Text betreffen. (Meta-)Daten bilden das Ergebnis des Annotationsprozesses. Allerdings wird der Terminus Annotation nicht nur für den Prozess der Informationsanreicherung selbst verwendet, sondern fallweise auch für das Ergebnis der Annotation, also für die hinzugefügten Informationen.

Der Vorteil digital annotierter Corpora besteht darin, dass sie durch spezielle Suchanfragen schnell und gezielt nach gleichartigen Informationen, die an verschiedenen Stellen verschiedener Texte zu finden sind, durchsucht werden können.

Digitales Annotieren, gleich, welche Informationen dadurch zu einem Text hinzugefügt werden, erfolgt idealerweise in mehreren Teilschritten: zuerst automatisiert (computergestützt) und dann semiautomatisiert (manuell durch den menschlichen Annotator/die Annotatorin am Computer zwecks Korrektur und zwecks Ergänzung der automatischen Annotationsergebnisse); beide Schritte unterliegen in der Regel standardisierten Verfahren. In der Regel greift man dabei auf sogenannte *Tagsets* (Annotationsinventare) zurück. Über solche Annotationsinventare wird weiter unten ausführlicher zu handeln sein. Zusätzlich liegen in der Regel noch umfangreichere Editionsrichtlinien (Guidelines) zu Grunde, die

kann, zu den Metadaten zählt.

⁸¹ Mit dieser Umschreibung von Annotation, die *Scherer*, *Korpuslinguistik* (wie Anm. 30), 21 entnommen ist, ist jene bei *Keibel*, *Kupietz*, *Perkuhn*, *Korpuslinguistik* (wie Anm. 58), 57f. inkompatibel. Letztere zählen das Thema des Textes noch zu den Metadaten (!) auf Textebene, ebenso seinen Umfang sowie seine Referenzen auf andere Texte. Thema und Umfang eines Textes jedoch unter den Oberbegriff der Metadaten zu subsumieren, halte ich für reichlich verfehlt, da beides aus einem Text selbst explizit hervorgeht oder in der Regel wenigstens hervorgehen sollte. Im Falle der Referenzen eines Textes zu anderen Werken ergibt sich das Problem, dass intertextuelle Bezüge sich unterschiedlichen Lesern je nach deren Kenntnisstand und Bildungsgrad jeweils unterschiedlich stark oder auch gar nicht erschließen. Insofern ist es nicht zwingend notwendig, intertextuelle Bezüge per se durch Metadaten explizit zu machen oder überhaupt als solche zu klassifizieren. Ganz abgesehen davon fällt die Kennzeichnung solcher Bezüge gar nicht in den Bereich der linguistischen Annotation, sondern geht darüber hinaus. Will man Referenzen eines Textes zu einem anderen als solche kenntlich machen, so erscheint es angebracht, nicht mehr von linguistischer Annotation zu sprechen, sondern von literarhistorisch- bzw. philologischer.

projektspezifisch formuliert werden. Tagsets und Editionsrichtlinien machen sowohl den Annotationsprozess als auch die Ergebnisse transparent und nachvollziehbar. Die einmal generierten Daten können, wie gesagt, wiederverwendet und für unterschiedlichste Fragestellungen herangezogen werden.

1.2.6.1) Tokenisierung, Tagging, Lemmatisierung

Der linguistischen Annotierung eines Textes geht in der Regel dessen Zerlegung (Segmentierung) in seine Einzelbestandteile (Tokens) voraus (Tokenisierung)⁸² bzw. eine wie auch immer geartete Explizierung seiner Teilelemente für den Computer. In welche Bestandteile, Segmente bzw. Tokens ein Text aufgeteilt wird, ist dabei davon abhängig, auf welcher Ebene man annotieren will und wie (mit welchen digitalen Werkzeugen, mit welchen Tagsets et c.). In einem ersten Schritt kann die Annotierung zum Beispiel den kompositorischen Aufbau eines Textes verdeutlichen (strukturelle Annotierung). Auf dieser Ebene wird noch nicht auf die Syntax der Sprache fokussiert, sondern auf eine allenfalls vom Autor (oder auch von einem Editor) mehr oder weniger intendierte Binnenstruktur eines literarischen, semiliterarischen oder pragmatisch-archivalischen (Amts-)Textes. Das heißt: Versteht man den Terminus „Token“ möglichst offen und breit, so bezieht er sich in diesem Fall auf ganze Textstrecken. Allerdings wird der Begriff Token innerhalb der Corpuslinguistik nur selten in dieser Bedeutung verwendet. Im einfachsten Fall werden Einheiten wie Überschriften, Zwischenüberschriften, Fließtext allgemein, Kapitel, Unterkapitel, Paragraphen, Strophen, Verse, Zeilen et c. voneinander isoliert (tokenisiert) und mittels „tag“ als solche gekennzeichnet (annotiert). Dabei können Paare von zusammengehörigen Tags der selben Art Tokens auch beiderseits ummanteln, um sie als solche zu kennzeichnen, so etwa am Anfang und am Schluss von distinkten Textstrecken. Der Begriff der Ummantelung wird weiter unten nochmals genauer thematisiert.⁸³ Der englische

⁸² Unter dem Begriff Token ist allgemein eine sprachliche Einheit in einem Textcorpus zu verstehen. Es kann sich hierbei um einzelne Worte handeln, aber auch um ganze Phrasen, Sätze, Absätze, je nach dem, was man als Token definiert; vgl. *Scherer*, *Korpuslinguistik* (wie Anm. 30), 32 f.

⁸³ Die Annotation von ganzen Textstrecken erfolgt innerhalb der unterschiedlichsten Editions-wissenschaften meist in TEI-XML. Ob Einheiten wie Überschriften, Zwischenüberschriften, Seitenangaben oder Fließtext in jedem Fall vom Autor intendiert wurden, ob der Editor dergleichen Einteilungen vorzunehmen hat und wie sich vom Autor intendierte Textstruktur zur Arbeit des Editors generell verhält – dies alles stellt einen breiten Fragenkomplex innerhalb der

Begriff *tag* bedeutet Etikett, Namensschild; *to tag something* = etwas etikettieren, mit einem Namensschild versehen.⁸⁴ Je nach Textgattung und je nach speziellem Aufbau eines Textes bzw. je nach Forschungsinteresse kann die strukturelle Annotation ganz unterschiedliche Texteinheiten umfassen, so etwa Titel, Untertitel, Prolog/Einleitung, Hauptteile, Kapitel, Epilog, Fußnoten, (Inhalts-)Verzeichnisse, Register et c.; bei Urkunden würde man zumindest Protokoll, Kontext und Eschatokoll als solche annotieren, gegebenenfalls einschließlich aller Untereinheiten. Bei Briefen gilt es, Adresszeilen, Datums- und Ortsangaben (Wo und wann wurde der Brief abgefasst, wo und wann versendet?), Fließtext, Anrede- und Schlussgrußformeln, Unterschriften, Postscripta et c. voneinander zu unterscheiden und entsprechend zu markieren.

Von linguistischer Annotation im eigentlichen Sinne spricht man dann, wenn ein Text zuerst bis auf die Satz- beziehungsweise bis auf die Wort- und auf die Satzzeichenebene segmentiert (tokenisiert) wird, um Worte und Satzzeichen als einzelne Tokens entsprechend zu annotieren, das heißt: ihrer Art und ihren Eigenschaften nach zu kennzeichnen und mit dazugehörigen Informationen zu versehen.⁸⁵ Zusätzlich wird jeder Token, sofern es sich um einzelne Worte handelt

Editionswissenschaften dar, der hier bloß am Rande erwähnt, nicht jedoch näherhin ausgeführt werden soll.

⁸⁴ *Lemmitzer, Zinsmeister*, Korpuslinguistik (wie Anm. 53), 61 verstehen den englischen Begriff Tag allgemeiner – und zwar als Annotationskategorie schlechthin, sodass ein Tag nicht nur einem einzelnen Wort, sondern auch ganzen Sequenzen oder einer „Relation“ (ebd.) zugeordnet werden kann, wobei offen gelassen wird, was unter Relation konkret zu verstehen ist.

⁸⁵ Die Frage, was auf Wortebene als Token zu gelten habe, ist durchaus diffiziler, als es oberflächlich betrachtet erscheinen mag; vgl. *Jenset, McGillivray*, Quant. Hist. Ling. (wie Anm. 58), 111f. Die gängige graphematische Definition, wonach ein Wort-Token jene Zeichenfolge sei, die zwischen zwei Leerzeichen (*white spaces*) oder zwischen einem Leerzeichen und einem Interpunktionszeichen steht, erweist sich in der Praxis oftmals als zu begrenzt. Allein schon die Verwendung des Punktes bei Abkürzungen oder bei Datumsangaben kann Probleme mit sich bringen, zumal, wenn der Punkt je nach Sprache unterschiedlich verwendet wird – vgl. André *Hagenbruch*, Flache Satzverarbeitung. In: *Carstensen, Ebert et al.*, Computerlinguistik und Sprachtechnologie (wie Anm. 21), 264–279; hier: 264–266 & 278. Ebenso müssen kontrahierte Formen mit umgangssprachlichem Einschlag (zum, ans, hats) oder Mehrwortlexeme, Eigennamen und mehrteilige Toponyme (2-Liter-Kanister, Wiener Neustadt) immer wieder diskutiert und näher definiert werden, auch in Abhängigkeit von der jeweiligen Sprache: Gelten solche Zeichenketten nun als ein Token oder als mehrere [vgl. *Lemmitzer, Zinsmeister*, Korpuslinguistik (wie Anm. 53), 62]? Mitunter müssen mehrere Worte als ein Token zusammengefasst werden (z.B. New York). Ähnliche Fragen stellen sich, wenn man es mit älteren Sprachstufen und mit historischen Corpora zu tun hat, in denen die Gewohnheiten der Getrennt- und der Zusammenschreibung mitunter schwanken und vom heutigen Gebrauch abweichen können [vgl. *Czeitschner, Resch*, Austrian Baroque Corpus (wie Anm. 9), 46]. Je nach Sprache, je nach verwendetem Tagset und je nach editorischen Zielsetzungen wird man derartige Fragen jeweils unterschiedlich lösen; vgl. *Lemmitzer, Zinsmeister*, Korpuslinguistik (wie Anm. 53), 64 (einschließlich dortiger Anm. 16).

(„Wort-Tokens“), mit seinem standardisierten Grundwort bzw. einem Lexem oder Lemma in Verbindung gebracht (Lemmatisierung).⁸⁶ Dieser Vorgang geschieht meist unter Bezugnahme und im Abgleich mit ausgewählten Wörterbüchern, z.B. dem Duden für das moderne Standarddeutsch, dem Lexer für Mittelhochdeutsch oder dem Grimm-Wörterbuch für früheres Neuhochdeutsch.⁸⁷ Innerhalb der Corpuslinguistik wird der Begriff „Token“ meist für einzelne Worte bzw. Satzzeichen verwendet.

1.2.6.2) PoS-Tagging und morphologische Annotation

Erfolgen Segmentierung und Annotation auf Ebene der Worte und der Satzzeichen, fasst man also das einzelne Wort und das einzelne Satzzeichen als Token auf, so geschieht dies meist mit dem Ziel, die Wortarten und die morphologischen Merkmale der einzelnen Tokens durch Annotation explizit zu machen. Man spricht in diesem Fall von (Wortarten-)Tagging bzw. Part-of-Speech-(PoS-)Tagging,⁸⁸ in einem weiteren Schritt von morphologischer bzw. von morphologisch-syntaktischer Annotation.⁸⁹ Diese Arten der Annotation sind in der Corpuslinguistik sehr weit verbreitet, besonders das Wortarten-Tagging.

⁸⁶ Von den Begriffen Lemma bzw. Lexem und Token ist engl. *type* zu unterscheiden. Der Terminus Type fasst mehrere konkrete sprachliche Einheiten (z.B. Wort-Tokens) unter sich zusammen, die sich zwar durch Flexionsmerkmale, Schreib- oder Lautvarianten voneinander unterscheiden, aber alle zum selben kanonisierten Grundwort (Lemma) gehören, welches stellvertretend und übergeordnet für alle Wortformen steht, die in einem Corpus vorkommen. Die Types (Wortformen) Urteil, Urthel, Vrtel, Urtheil, Vrtheil und Urthl gehören als historische Schreib- bzw. Lautvarianten einschließlich aller abgewandelten Formen alle dem modernen Grundwort (Lemma) Urteil an; Beispiel entnommen aus: *Czeitschner, Resch, Austrian Baroque Corpus* (wie Anm. 9), 47f.; zur Unterscheidung von Tokens und Types vgl. auch *Scherer, Korpuslinguistik* (wie Anm. 30), 33 f.

⁸⁷ Im Falle historischer Sprachstufen kann mit dem Prozess der Lemmatisierung auf der Ebene der Tags eine Normalisierung historischer Grapheme einhergehen, das heißt: In der Annahme, dass man auch im Falle eines historischen Corpus eine digitale Suchanfrage gemäß moderner Schreibung durchführt, werden Grapheme in historischer Schreibung um ein Grundwort in moderner Standardschreibung angereichert; Bsp.: vgl. Anm. 86. Der Graphembestand des historischen Quelltextes selbst bleibt bei diesem Vorgang freilich unberührt und unverändert. Dieser Aspekt soll hier umso stärker hervorgehoben werden, als die Unversehrtheit des Quelltextes in der entsprechenden Passage bei *Lemnitzer, Zinsmeister, Korpuslinguistik* (wie Anm. 53), 85f. meines Erachtens etwas zu wenig betont wird. Der Graphembestand des Quelltextes bleibt unverändert, egal ob man sich bei der Lemmatisierung auf ein historisches oder ein modernsprachliches Wörterbuch bezieht! Die Lemmatisierung im Sinne einer (modernsprachlichen) Normalisierung bildet bloß eine „zusätzliche Annotationsebene“ (ebd.).

⁸⁸ Die Begriffe Part-of-speech-(PoS-)Tagging und Wortarten-Tagging sind ident.

⁸⁹ Auf den Unterschied zwischen reinem Wortarten-Tagging und der Annotation von Flexionsmorphologie wird in Kürze eingegangen.

Somit werden am einzelnen Wort Informationen über Wortart (*part of speech*), Wortgrundform (Lemma) sowie über Flexionsmerkmale des Wortes expliziert (z.B. Casus, Genus, Numerus, Person, Tempus und Modus). Es wird dabei meist von einer traditionellen Einteilung der Wortarten ausgegangen, wobei diese von Sprache zu Sprache und je nach zu Grunde liegender Grammatik-Theorie mehr oder weniger variieren können. Die explizite Zuordnung jedes Tokens zu seiner grammatikalischen Basiskategorie, seiner Wortart, stellt eine grundlegende Voraussetzung für folgende morphosyntaktische Analysen und Annotation dar,⁹⁰ worauf wir im weiteren Verlauf dieser Arbeit noch näher eingehen werden.

Wie zuvor bereits angedeutet, erfolgen das Wortarten-Tagging sowie die morphologische Annotation zumeist an Hand von Tagsets, also an Hand von Listen (Annotationsinventaren, -richtlinien bzw. -schemata oder -guidelines), welche für jeweils eine Wortart bzw. ein morphologisches Phänomen genau eine standardisierte Abkürzung als Namensschild (*label, tag*) vorsehen und die solche Abkürzungen idealiter für möglichst alle Wortarten, für möglichst alle morphologischen Phänomene und für möglichst alle Typen von Satzzeichen, die in einem Text vorkommen, enthalten. Jede einzelne dieser standardisierten Abkürzungen fungiert als *tag*, als „Namensschild“ oder Label. Bei der linguistischen Annotation wird in der Regel jeder Token eines Textes (das heißt: jedes Wort, jedes Satzzeichen) mit mindestens einem entsprechenden Tag versehen. Dieser macht die dem Token eigenen Charakteristika explizit, z.B. seine Wortart (Verb, Nomen, Adjektiv, Adverb o.ä.). Wortarten-Tags werden besonders häufig verwendet. Die zusätzliche Kennzeichnung von (Ab-)Sätzen, Satzgefügen, von größeren kompositorischen Strukturen und von längeren Wortgruppen als Tokens im weitesten Sinne ist nicht immer üblich. Als *das* Standard-Inventar von Tags (Abkürzungen) hat sich für deutschsprachige Corpora de facto das sogenannte Stuttgart-Tübingen-Tagset (STTS) etabliert, das in der zweiten Hälfte der 1990er-Jahre in Kooperation von Wissenschaftler_innen-Teams der dortigen Universitäten entwickelt wurde.⁹¹ Es kommt zumeist im Falle modernsprachlicher Corpora zum

⁹⁰ Vgl. Michael Hess, Tagging (Zürich 2006), 2; online unter: <https://files.ifi.uzh.ch/cl/hess/classes/le/tag.0.1.pdf>; letzter Zugriff: 30.05.2019.

⁹¹ Vgl. <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>; letzter Zugriff: 30.05.2019; eine Vorstufe: <http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1995.pdf>; letzter Zugriff: 30.05.2019.

Einsatz, wird aber auch für historische Textsammlungen mit älteren Sprachstufen des Deutschen verwendet und adaptiert,⁹² obwohl für Letztere bereits eigene Tagsets entwickelt wurden.⁹³ Im STTS steht etwa, um nur ein Beispiel zu nennen, die Abkürzung *VFIN* für finites Vollverb, z.B. für den Token „schreibt“ (im Gegensatz zu Modal- und Auxiliärverben).⁹⁴

Arbeitet man nun, was den Normalfall darstellt, am Computer mit einem digital vorliegenden Textcorpus und annotiert dieses linguistisch mittels STTS, so wird, wie bereits erwähnt, jeder einzelne (Wort- und Satzzeichen-)Token des Textes seiner Art und seinen Eigenschaften nach mit dem jeweils dazugehörigen Tag, also mit einer standardisierten Abkürzung nach STTS, versehen. Darin besteht dem Grundgedanken nach linguistische Annotation.

Zu betonen ist in diesem Zusammenhang, dass ein Tagset „immer einen Kompromiss [...] zwischen Genauigkeit und Handhabbarkeit [darstellt]“⁹⁵. Kaum ein Tagset weist in der Praxis die Kapazitäten auf, um tatsächlich *alle* Wort- und Verwendungsarten sowie alle syntaktischen Funktionen, wie sie für eine sprachliche Zeichenfolge gebräuchlich und möglich sind, eindeutig darzustellen, vor allem bei älteren und bei weniger standardisierten Sprachstufen; dies vor allem deshalb, weil die meisten Tagsets darauf ausgerichtet sind, zunächst als Vorlage und als Matrix für die ersten Schritte der Annotation eines Textes nach der Segmentierung zu dienen, nämlich für die automatisierte (Wortarten-)Annotation mittels Computer-Tool, also mittels Tagger. Man unterscheidet grundsätzlich zwischen stochastischen, regelbasierten und transformationsbasierten Taggern.⁹⁶ Zweitere sind häufig auf eine einzige Sprache ausgerichtet und wurden in den letzten Jahren in der Regel von statistischen Taggern ersetzt, die sprachunabhängig arbeiten.⁹⁷

⁹² So etwa im Falle des Austrian Baroque Corpus (ABaC:us); vgl. *Czeitschner, Resch*, Austrian Baroque Corpus (wie Anm. 9), 46ff.

⁹³ Vgl. *Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, Klaus-Peter Wegera*, HiTS: ein Tagset für historische Sprachstufen des Deutschen. In: *Journal for Language Technology and Computational Linguistics – JLCL* 28/1 (2013), 85–137.

⁹⁴ Die volle Liste an Abkürzungen (Tags) für alle Wortarten und Satzzeichen nach STTS findet sich in: <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>, 6 (entspricht 7 in der Zählung der pdf-Datei); Zugriff: 30.05.2019.

⁹⁵ *Lemnitzer, Zinsmeister*, Korpuslinguistik (wie Anm. 53), 65.

⁹⁶ Zur Funktionsweise und zur Verlässlichkeit einiger gängiger PoS-Tagger vgl. näherhin *Bubenhof, Konopka, Schneider*, Präliminarien (wie Anm. 66), 149–181; ferner *McGillivray*, *Methods in Latin Computational Linguistics* (wie Anm. 50), 22 f.

⁹⁷ Vgl. *McGillivray*, *Methods in Latin Computational Linguistics* (wie Anm. 50), 22.

Computer-Tools benötigen in vielen Fällen ein *begrenzt*es und eindeutiges Repertoire an Regeln und Möglichkeiten, um auf die automatische Erkennung sprachlicher Phänomene hin trainiert zu werden. Hier zeigt sich immer wieder, dass Tagger an ihre Grenzen stoßen, wenn es etwa darum geht, ambige Worte ihrer Verwendung nach eindeutig einer bestimmten Wortart zuzuordnen, sodass für eine möglichst fehlerfreie Annotation nach dem ersten Schritt der automatisierten (Vor-)Verarbeitung manuelle Kontrollen notwendig werden, nämlich Kontrollen der automatischen (Wortarten-)Zuordnungen und der Lemmatisierung (zu diesem Begriff etwas weiter unten).⁹⁸

Nimmt man beispielsweise den Satz: „Das scheint sein Koffer zu sein“, so tritt der Token „sein“ einmal als Besitz anzeigendes Pronomen und einmal als Infinitiv des Auxiliärverbs *sein* auf. Ungeachtet der syntaktischen Umgebung kann die Zeichenfolge „sein“ demnach mit zwei unterschiedlichen Tags versehen werden. Derartige Ambiguitäten stellen automatische Annotationstools potenziell vor Probleme und bedürfen von Fall zu Fall sorgsamer Nachkontrolle. Besonders dann, wenn man mit historischen Textcorpora und älteren Sprachstufen arbeitet, bleiben die Leistungen mancher Tagger trotz technologischer Fortschritte weit hinter den Erwartungen und hinter den gewünschten Ergebnissen zurück.⁹⁹ Man unterscheidet streng genommen zwischen reinem Wortarten-Tagging [engl.: *Part of Speech- (PoS-)Tagging*] und morphologischer Annotation. Wortarten-Tagging geht Letzterer mitunter voraus oder sie läuft parallel zu ihr. Sie zieht morphologische Annotation aber nicht zwangsläufig nach sich. PoS-Tagging kann auch unabhängig von morphologischer Annotation erfolgen. Aber: Nur bei morphologischer Annotation werden auch Flexionsmerkmale von Worten berücksichtigt und gekennzeichnet.

Nach Segmentierung eines Textes in Tokens, ein Prozess, der im Vorfeld linguistischer Annotation eine zentrale Rolle spielt, wird beim Wortarten-Tagging, wie der Begriff schon zeigt, jedem Wort-Token ein Tag zugewiesen, der dessen Wortart bezeichnet. Auch im Falle des PoS-Taggings erfolgt dieser Vorgang in der Regel zuerst automatisiert durch einen entsprechenden Tagger und sollte in einem

⁹⁸ Weitere Beispiele für ambige, also mehrdeutige Wortformen, die bei der computerbasierten Vorverarbeitung entsprechend Probleme bereiten können: vgl. *Lemnitzer, Zinsmeister*, Korpuslinguistik (wie Anm. 53), 65 f.

⁹⁹ Vgl. *Czeitschner, Resch*, Austrian Baroque Corpus (wie Anm. 9), 45.

zweiten Schritt manuell kontrolliert werden.¹⁰⁰ Bei der morphologischen Annotation schließt an Segmentierung des Textes in Tokens und an das Tagging der Wortarten zusätzlich die Anreicherung mit morphologischen Informationen auf Wortebene an. Zusätzlich wird jeder Token mit seinem Grundwort, seinem Lemma, „wie es im Wörterbuch steht“, versehen (Lemmatisierung).

Morphologische Annotation kann Wortarten-Annotation also beinhalten, bleibt bei dieser jedoch nicht stehen. Sie geht darüber hinaus, indem sie den einzelnen Wort-Token nicht nur mit Informationen zu seiner Wortart, sondern auch zu seiner Morphologie anreichert, also, wie gesagt, mit Informationen etwa über Casus, Genus, Numerus, Person, Tempus, Modus et cetera. Allerdings: Den Terminus der morphologischen Annotation findet man in der Literatur mitunter als Überbegriff verwendet, unter den das PoS-Tagging fallweise subsumiert wird¹⁰¹, doch kommt auch der umgekehrte Fall vor (PoS-Tagging als Überbegriff für Wortarten-Annotation *und* für morphologische Annotation gleichzeitig).¹⁰² Unabhängig von linguistischen Interferenzen zwischen Morphé, Syntax und Wortart plädieren wir an dieser Stelle dringend für eine terminologisch saubere Trennung zwischen morphologischer Annotation einerseits und PoS-Tagging andererseits, da es sich in der Annotationspraxis um zwei unterschiedliche Vorgänge handeln *kann*, die getrennt voneinander durchgeführt werden *können* und einander überhaupt nicht bedingen, weder in die eine noch in die andere Richtung.

¹⁰⁰ Die hohe Treffergenauigkeit von automatischen Taggern im Falle des PoS-Taggings, die *Lemmitzer, Zinsmeister*, *Korpuslinguistik* (wie Anm. 53), 68 (einschließlich dortiger Anm. 27) ins Treffen führen [95 bis 98 % Wort-Akkuratheit (!)] sollte, wenn überhaupt, nur für hoch- und für modernsprachliche Corpora mit einem sehr hohen Grad an sprachlicher Standardisierung veranschlagt werden. Für historische Textsammlungen mit älteren Sprachstufen und mit entsprechend geringerem Ausmaß an sprachlicher Standardisierung ist eine ähnlich hohe Trefferquote jedoch völlig unrealistisch. Ebenso wenig lässt sich eine genaue Trefferquote von automatisierten Taggern derzeit exakt oder allgemein beziffern, wenn man die Tools auf historische Corpora anwendet, da die Ergebnisse im Moment von Corpus zu Corpus noch stark variieren können [vgl. *Czeitschner, Resch*, *Austrian Baroque Corpus* (wie Anm. 9), 45f. (einschließlich dortiger Anm. 13)].

Ganz abgesehen davon scheinen *Lemmitzer, Zinsmeister* (ebd.) den Begriff Tagging ausschließlich für das automatisierte (!) PoS-Tagging zu verwenden, was im Kontext ihrer Gesamtpublikation aber nicht konsequent durchgehalten wird und im Kontext der linguistischen Annotation ganz generell auch nicht konsequent durchzuhalten ist. Oberflächlich betrachtet entsprechen die Begriffe Annotation und Tagging einander sogar weitestgehend.

¹⁰¹ Vgl. *Jenset, McGillivray*, *Quant. Hist. Ling.* (wie Anm. 58), 112 ff.

¹⁰² Vgl. Erhard *Hinrichs*, Tylman *Uhle*, *Linguistische Annotation*. In: *Lemmitzer, Lobin*, *Texttechnologie. Perspektiven und Anwendungen* (wie Anm. 70), 217–243, speziell: 219.

1.2.6.3) Treebanking und Parsing

Nur kurz sei an dieser Stelle auf die Möglichkeit hingewiesen, Wort- und Satzzeichen-Tokens auch in Hinblick auf ihre syntaktische Funktion und in Hinblick auf ihre Beziehungen zu anderen Tokens im Satz zu annotieren (syntaktische Annotation/*parsing* – zu diesem Begriff etwas weiter unten). Diese Art der Annotation bringt es häufig mit sich, dass die Beziehungen der einzelnen Tokens zueinander graphisch dargestellt werden, unter anderem in Form von sogenannten Baumbanken (*treebanks*):¹⁰³ „A treebank is a large collection of sentences that have been syntactically annotated.“¹⁰⁴

Diese höchst aufwendige, weil fast nur manuell durchführbare Methode, die Erstellung solcher Baumbanken („*treebanking*“), hat innerhalb der computer- und innerhalb der corpuslinguistischen Forschung der letzten Jahre sehr viel Aufmerksamkeit erfahren und sehr hohe Reputation genossen, da sie ihrem Prinzip nach die umfassendste Basis für digital-linguistische Anfragen an annotierte Corpora darstellt. Im Falle lateinischer Texte sind gegenwärtig die Ergebnisse mindestens dreier größerer Projekte verfügbar, die die Erstellung von Treebanks zum Ziel hatten: der Index Thomisticus Treebank, der Latin Dependency Treebank (assoziiert mit der Perseus Digital Library) und der PROIEL-Treebank.¹⁰⁵ Im Gegensatz zur morphologischen Annotation auf Token-Ebene wird bei der syntaktischen Annotation und beim Treebanking auf die Beziehung mehrerer Tokens zueinander fokussiert, also auf deren syntaktisch bedingte „Verbindungen“ (syntagmatische Relationen¹⁰⁶). Graphisch dargestellt können diese Verbindungen ein baumartiges Geflecht zwischen den einzelnen Bestandteilen (Tokens) eines Satzes ergeben,¹⁰⁷ woher der Name der Baumbanken

¹⁰³ Vgl. *Jenset, McGillivray*, Quant. Hist. Ling. (wie Anm. 58), 115 ff. und *Lemnitzer, Zinsmeister*, Korpuslinguistik (wie Anm. 53), 71 ff.

¹⁰⁴ David *Bamman*, Gregory *Crane*, Building a Dynamic Lexicon form a Digital Library. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (New York 2008), 11–20; hier: 12; online unter: <http://people.ischool.berkeley.edu/~dbamman/pubs/pdf/jcdl2008.pdf>; letzter Zugriff: 30.05.2019.

¹⁰⁵ Details zusammengefasst bei *McGillivray*, Methods in Latin Computational Linguistics (wie Anm. 50), 26f.; die mehr oder weniger große Anzahl annotierter sprachlicher Tokens in diesen Projekten soll nicht darüber hinwegtäuschen, dass die Treebanks linguistisch bisher kaum ausgewertet wurden.

¹⁰⁶ Vgl. *Lobin*, Computerlinguistik und Texttechnologie (wie Anm. 28), 23–24.

¹⁰⁷ Zur Veranschaulichung vgl. e.g. <https://digitalfellows.commons.gc.cuny.edu/2015/10/08/treebanking-with-arethusa/>;

bzw. der Treebanks sich ableitet. Daneben gibt es etliche weitere Möglichkeiten, syntaktisch annotierte Tokens bzw. deren Beziehungen untereinander graphisch darzustellen, beispielsweise in Form einer Klammerstruktur, in Form einer funktionalen Dependenzstruktur, sowie hybride Modelle, die in der zitierten Sekundärliteratur allesamt zusammenfassend visualisiert werden.¹⁰⁸

Im Gegensatz zum leichter standardisierbaren und daher eher konventionalisierten PoS- und zum morphologischen Tagging ist das graphische Ergebnis syntaktischer Annotation in größerem Maße davon abhängig, welche linguistischen Theorien der Annotationspraxis zu Grunde liegen.¹⁰⁹

In Hinblick auf die hier skizzierten Möglichkeiten, syntaktische Beziehungen zwischen einzelnen Tokens graphisch darzustellen, sind aus unserer Sicht zwei Ergänzungen vorzunehmen; zum einen, dass dergleichen graphische Repräsentationen vor allem zwei Chancen bieten: Sie können einerseits für linguistische Fragestellungen von Belang sein, beispielsweise für Vergleiche syntaktischer Strukturen mehrerer, sprachgeschichtlich auch nicht verwandter Sprachen, andererseits für didaktische Zwecke, um etwa im schulischen Kontext die strukturellen Zusammenhänge innerhalb von Sätzen zu verdeutlichen. Letzteres allerdings erfordert eine nicht geringe Emanzipation vom traditionellen Verständnis von Grammatik, aus deren Sicht die hierarchische Strukturierung von Sätzen, wie sie für syntaktische Annotation nötig ist, ungewohnt erscheint, sieht man einmal von der üblichen Unterteilung in Haupt- und Nebensätze ab. Treebanking erfordert in den allermeisten Fällen eine profunde Einarbeitung in die Funktionsweise von Dependenzgrammatiken, welche nicht näherhin Gegenstand der vorliegenden Untersuchung sein sollen.

Zum anderen ist zu ergänzen, dass dergleichen graphische Repräsentationen syntaktischer Beziehungen nicht zwangsläufig in den Kontext digitaler Editionen gehören müssen. Letztere sollten als Endprodukt editorischer und digitaler Arbeit eine im World-Wide-Web verfügbare, optisch ansprechende und übersichtliche graphische Benutzeroberfläche für den User bieten, über die all ihre Teileigen-

letzter Zugriff: 30.05.2019.

¹⁰⁸ Vgl. *Lemnitzer, Zinsmeister*, *Korpuslinguistik* (wie Anm. 53), 71–80.

¹⁰⁹ Vgl. *Hinrichs, Uhle*, *Linguistische Annotation* (wie Anm. 102), 221 ff.

schaften leicht zugänglich sind.¹¹⁰ Graphische Repräsentationen syntaktisch annotierter Sätze können im Kontext digitaler Editionen allenfalls ein zusätzliches Angebot unter mehreren visuellen Features mit ganz unterschiedlichen Funktionen darstellen. Im Kontext gedruckter Editionen sind Treebanks, abgesehen von einzelnen Beispielen, überhaupt undenkbar, da sie, auf das Papier projiziert, sehr viel Platz benötigen. Ein einzelner Satz könnte, als Baumstruktur dargestellt, eine ganze Seite in Anspruch nehmen oder überhaupt mehrere.

Ganz generell ist die graphische Darstellung von annotierten oder zu annotierenden Texten *innerhalb ihrer Annotationswerkzeuge* streng zu unterscheiden von der graphischen Darstellung des „fertig edierten“ Textes innerhalb einer (digitalen) Edition.

Im Bereich der syntaktischen Annotation ist abschließend der Begriff *Parsing* zu erläutern.¹¹¹ Seitens der gängigen Literatur wird dieser Terminus mitunter nur sehr ungenau umrissen und nicht alle Definitionen bzw. Definitionsversuche sind brauchbar. Richtig, doch gleichwohl sehr allgemein gehalten sind folgende: „Das P. bezeichnet allgemein den Prozess der syntaktischen Textanalyse.“¹¹² „Parsing consists in automatically [!] annotating a corpus from a syntactic point of view.“¹¹³ Auf der anderen Seite scheint es Definitionen zu geben, die anderen widersprechen, je nach dem, aus der Perspektive welcher linguistischen Teildisziplin Parsing beschrieben wird. Innerhalb der Computerlinguistik hat Parsing sich als so „selbstverständlicher Grundbegriff“¹¹⁴ etabliert, dass dieser einer eingehenderen Erklärung nicht immer für wert befunden wird.

Grundsätzlich handelt es sich beim Parsing um digital unterstützte Satzanalyse, die der syntaktischen Annotation voraus- bzw. mit ihr Hand in Hand geht. Die entsprechenden elektronischen Werkzeuge, Parser, überprüfen, ob ein Satz einer bestimmten vor- bzw. übergeordneten Grammatik entspricht und ihr deshalb zugeordnet werden kann. Parser „analysieren die Struktur von Sätzen und fügen

¹¹⁰ Hierzu ausführlicher: Patrick Sahle, Digitale Editionen. In: Jannidis, Kohle, Rehbein, Digital Humanities (wie Anm. 22), 234–249.

¹¹¹ Vgl. Lobin, Computerlinguistik und Texttechnologie (wie Anm. 28), 41–53.

¹¹² Lemnitzer, Zinsmeister, Korpuslinguistik (wie Anm. 53), 198.

¹¹³ Jensen, McGillivray, Quant. Hist. Ling. (wie Anm. 58), 118.

¹¹⁴ Sven Naumann, Hagen Langer, Parsing. Eine Einführung in die maschinelle Analyse natürlicher Sprache (Trier 1994), 4; online unter: <https://www.uni-trier.de/fileadmin/fb2/LDV/Naumann/BUCH1-1.pdf>; letzter Zugriff: 30.05.2019.

Informationen auf Satzebene in den Text ein. Sie bestimmen [...] syntaktische Kategorien und syntaktische Funktionen“¹¹⁵, so etwa in Bezug auf einzelne Satzteile und im Unterschied zum PoS-Tagging: Nominalphrase, Verbalphrase, Präpositionalphrase, Subjekt, Prädikat et c., Größen also, welche, wie oben skizziert, in ihrer Beziehung zueinander graphisch dargestellt werden können. „Ziel des Parsings ist es, [...] Verfahren [...] zu beschreiben, die es erlauben, [bezüglich eines Satzes] festzustellen, ob dieser Satz durch [eine bestimmte] Grammatik beschrieben wird und somit zu [einer bestimmten] Sprache gehört oder nicht [...].“¹¹⁶ Wie im Falle der Tagger gibt es auch unter den Parsern solche, die statistisch und solche die regelbasiert arbeiten.¹¹⁷

In der Computerlinguistik wird Parsing häufig „mit automatischer (syntaktischer) Satzanalyse gleichgesetzt.“¹¹⁸ Die Unschärfen des Begriffs scheinen auch darin begründet zu sein, dass der Terminus Parsing innerhalb der Computerlinguistik bzw. innerhalb verwandter sprachwissenschaftlicher Disziplinen im Laufe der Zeit eine Bedeutungsverschiebung erfahren hat: Parsing meinte ursprünglich die Wortarten-, also die PoS-Annotation (!), wofür heute der Begriff Tagging gängig ist.¹¹⁹

¹¹⁵ Scherer, *Korpuslinguistik* (wie Anm. 30), 72.

¹¹⁶ Helmut Schmidt, *Parsing I. Seminarskript* (Stuttgart 2003), 3; online unter: <http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schmid/ParsingI/parsing.pdf>; letzter Zugriff: 04.07.2019.

¹¹⁷ Vgl. McGillivray, *Methods in Latin Computational Linguistics* (wie Anm. 50), 27.

¹¹⁸ Naumann, Langer, *Parsing* (wie Anm. 114), 4.

¹¹⁹ Vgl. Hagen Langer, *Syntax und Parsing*. In: Carstensen, Ebert et al., *Computerlinguistik und Sprachtechnologie* (wie Anm. 21), 280–329; hier: 303.

1.2.6.4) Rekapitulation

Damit sind die wichtigsten linguistischen Annotationsarten ihrem Wesen nach in Grundzügen dargestellt, ebenso die wichtigsten Arbeitsschritte im Annotationsprozess. Diese Arbeitsschritte seien hier noch einmal zusammenfassend rekapituliert: erstens Segmentierung des Textes in einzelne Tokens (z.B. auf Satz-, Satzzeichen- und auf Wortebene), zweitens Anreicherung der Tokens mit Informationen durch (und in) Tags (meistens mit Wortarten-Tags und Tags, die morphologische Informationen explizit machen), drittens Anreicherung der Wort-Tokens mit ihrem Grundwort (Lemma > Lemmatisierung). Darin, in diesen drei Schritten, besteht dem Grundgedanken nach linguistische Annotation¹²⁰ und sofern sie, was den Normalfall darstellt, computergestützt erfolgt, sprechen wir von digital-linguistischer Annotation. Sie wird innerhalb der Corpus- und der Computerlinguistik, innerhalb der Texttechnologie und innerhalb der Digital Humanities, schließlich innerhalb des digitalen Editionswesens im Speziellen, gleichermaßen angewandt.

1.2.6.5) Semantisch-pragmatische Annotation

Wir wollen es im Rahmen der vorliegenden Arbeit bei dieser groben Darstellung belassen. Hinsichtlich weiterer Annotationsarten, speziell in Bezug auf semantische und auf pragmatische Annotation, „indicating the semantic fields of a text“¹²¹, sei an dieser Stelle einmal mehr auf die bereits zitierte Literatur verwiesen.¹²² Zur Veranschaulichung soll in diesem Zusammenhang speziell eine österreichische Publikation hervorgehoben werden, die als Vorarbeit zum Austrian Baroque Corpus im Jahr 2011 entstanden ist, bei der verschiedene Umschreibungen der Entität „Tod“ bzw. von „sterben“ semantisch annotiert wurden (codiert in Form von TEI-XML; zu diesem Begriff etwas weiter unten): „Although personifications of violent death were very popular at that time [i.e. in the 17th century], the texts also allow other conceptual representations of death and dying. A primary aim of the project was thus to apply semantic annotations

¹²⁰ Vgl. auch *Lobin*, Computerlinguistik und Texttechnologie (wie Anm. 28), 99–102.

¹²¹ *Jenset, McGillivray*, Quant. Hist. Ling. (wie Anm. 58), 119.

¹²² Vgl. *Scherer*, Korpuslinguistik (wie Anm. 30), 22 & 59; *Lemnitzer, Zinsmeister*, Korpuslinguistik (wie Anm. 53), 81–87; *Jenset, McGillivray*, Quant. Hist. Ling. (wie Anm. 58), 119–122.

for rendering such conceptual diversity more easily discernible and to support access to comparable digital resources. [...] Instances of death as a personified entity are marked-up in the following manner:

```
<rs type="death" subtype="figure">Mors, Tod, Todt</rs>  
<rs type="death" subtype="figureAlternative">General Haut und Bein,  
Menschenfeind</rs>  
<rs type="death" subtype="attribute">knochenreich, ohnartig,  
vnersättlich</rs> [...]  
<rs type="death" subtype="event">den Geist aufgeben, aus der Welt  
schleichen</rs>“123.
```

Kurz erwähnt seien darüber hinaus die soziolinguistische sowie die diskurs- bzw. die textlinguistische Annotation und die problemorientierte Annotation (Fehlerannotation):

1.2.6.6) Soziolinguistische Annotation

Soziolinguistisch annotierte Corpora können kontextuelle Informationen zu jenen Personen enthalten, die in ihnen genannt werden (z.B. deren Alter, ihre Herkunft, ihren sozialen Stand, ihre Muttersprache, ihren Beruf, ihr religiöse Bekenntnis et c.). Sie können Informationen zu den verwandtschaftlichen oder zu den zwischenmenschlichen Beziehungen mehrerer Personen untereinander enthalten sowie (etwa im Falle von Briefen) Angaben zum Grund ihrer gegenseitigen Kommunikation. Dabei wird, wie oben ausgeführt, im Sinne der Soziolinguistik davon ausgegangen, dass diese Parameter und die verwendete Sprache einander beeinflussen.

¹²³ Ulrike Czeitschner, Thierry Declerck, Karlheinz Moerth, Claudia Resch, Gerhard Budin, A Text Technology Infrastructure for Annotating Corpora in the eHumanities. In: Francesca Borri, Stefan Gradmann, Carlo Meghini, Heiko Schuldt (Hgg.), Research and Advanced Technology for Digital Libraries. Proceedings of the International Conference on Theory and Practice of Digital Libraries (Berlin/Heidelberg 2001), 457–460; hier: 458.

1.2.6.7) Diskurs-/Textlinguistische Annotation

Die diskurs- bzw. die textlinguistische Annotation erfasst unter anderem „Phänomene wie die sprachliche Markierung von Höflichkeit“¹²⁴. Dies kann dann von Belang sein, wenn etwa danach gefragt wird, wie gegenseitige Anreden und Anredeformeln vor allem zu Beginn und am Schluss von Gelehrtenbriefen gelehrte Beziehung und Hierarchien unter den Korrespondenzpartner inszenieren, bestätigen, verdeutlichen oder in Frage stellen. Über diskurs- bzw. die textlinguistische Annotation lässt sich eine Verbindung zwischen Corpuslinguistik und (historischer) Soziolinguistik herstellen.

1.2.6.8) Fehlerannotation

Abschließend erscheint noch die sogenannte problemorientierte Annotation bzw. Fehlerannotation von Bedeutung. Diese Annotationsart wird oftmals im Falle von Corpora angewandt, die den meist schriftlichen Sprachgebrauch beim Erlernen einer Fremdsprache dokumentieren, also im Falle von sogenannten Lern(er*innen)corpora. Hier konzentriert man sich vor allem darauf, Phänomene zu annotieren, die vom regelkonformen Standard der zu erlernenden Ziel(schrift)sprache abweichen („Fehler“).¹²⁵

Dieser Ansatz erscheint für die Belange der vorliegenden Arbeit *mutatis mutandis* vor allem aus zwei Gründen von Interesse – erstens, weil die lateinische Sprache mit den fundamentalen kulturgeschichtlichen Umwälzungen der Spätantike (des Frühmittelalters) als Muttersprache ausstarb und sukzessive für alle, die sie gebrauchten, zur Fremd- bzw. zur Zweitsprache wurde, die erst erlernt werden wollte; zweitens, weil man gerade innerhalb der frühneuzeitlichen *Res publica literaria* Wert darauf legte, sich wenigstens theoretisch eines Sprachgebrauchs zu befleißigen, der sich beträchtlich vom kirchlich geprägten Latein im Mittelalter unterschied und der sich hinsichtlich Syntax und Semantik bekanntlich wieder eng an der als modellhaft empfundenen Sprache der „klassischen“ Autoren orientierte, vor allem an jener Ciceros, aber auch an Ovid, Sallust, Seneca und Vergil.

¹²⁴ Scherer, Korpuslinguistik (wie Anm. 30), 22.

¹²⁵ Vgl. Scherer, Korpuslinguistik (wie Anm. 30), 22; detaillierter: Lemnitzer, Zinsmeister, Korpuslinguistik (wie Anm. 53), 85f.

Vor dem Hintergrund, dass man den Gebrauch des mittelalterlich-kirchlichen Lateins unter frühneuzeitlichen Gelehrten als verpönt ansah, wirkt es gerade besonders interessant, Gelehrtenbriefe auf sprachliche Phänomene zu untersuchen und zu annotieren, die vom klassischen Standard-Latein abweichen, die aus klassischer Perspektive gleichsam „Fehler“ darstellen und die entgegen der klassischen Norm Residuen mittelalterlichen Sprachgebrauchs durchschimmern lassen.¹²⁶

1.2.6.9) Codierung und XML

Wir wollen nun noch einmal auf den Aspekt der graphisch-visuellen Darstellung von Texten eingehen, die annotiert werden sollen.¹²⁷ Grundsätzlich gilt es hierbei zu berücksichtigen, dass Annotation für gewöhnlich digital vorgenommen wird, also mittels Computer und „auf diesem“ mit entsprechendem Programm, Tool oder Werkzeug, wobei zwischen den letztgenannten drei Ausdrücken Programm – Tool – Werkzeug hier kein terminologischer Unterschied gemacht werden soll. Zwischen Mensch und Maschine ergibt sich dadurch folgende Diskrepanz: Der menschliche Bearbeiter/die Bearbeiterin verfügt kraft seiner/ihrer Inkulturation bereits zu Beginn der Arbeit über große Mengen impliziten Wissens in Bezug darauf, woraus ein Text besteht, wie er strukturiert ist, etwa, wo ein Wort, wo ein Satz, wo eine Zeile endet, wo eine neue beginnt, wo ein Absatz anfängt und durch welche optischen Marker dergleichen Textbestandteile voneinander abzugrenzen sind.

Im Vorfeld digitaler Annotation können wir aber nicht davon ausgehen, dass der Computer kraft seiner Programmierung über dieselbe Menge „impliziten Wissens“ zur Struktur eines Textes verfügt wie wir als dessen menschliche Bearbeiter*innen. Ohne zusätzliche Strukturierung stellt ein Text für einen Computer zunächst nichts anderes dar, als einen weitestgehend unstrukturierten Zeichenstrom – Rohdaten also. Daraus ergeben sich zweierlei Notwendigkeiten: Einerseits muss der Text so strukturiert werden, dass seine voneinander

¹²⁶ Auf der Ebene der Lexik werden solche Phänomene vor allem dort zu erwarten sein, wo kirchliche Themen besprochen werden und die Korrespondenzpartner nicht umhin können, lateinisches Vokabular kirchlichen Gepräges zu verwenden.

¹²⁷ Zum folgenden vgl. u.a. *Jenset, McGillivray*, Quant. Hist. Ling. (wie Anm. 58), 103–106 und: *Hinrichs, Uhle*, Linguistische Annotation (wie Anm. 102), 225–231.

abgegrenzten Grundbestandteile auch für den Computer als solche erkennbar, verarbeitbar und speicherbar sind – auf Basis strukturierender Metadaten. Andererseits sollte diese Strukturierung in einer Art und Weise erfolgen, dass auch der menschliche Bearbeiter/die Bearbeiterin mit dem so strukturierten Text nach wie vor arbeiten kann. Der Text muss so *codiert* sein dass beiden Aspekten Rechnung getragen wird. Allerdings muss der Bearbeiter/die Bearbeiterin „nicht notwendigerweise mit dem Dateiformat [selbst] in Berührung kommen, denn meist [orientiert sich] die Darstellung am Bildschirm [...] unabhängig vom Dateiformat [...] primär an den Bedürfnissen des Betrachters.“¹²⁸

Zum Zweck der Annotation ist etwa die vertikale Aufsplittung und Darstellung von Texten in Form von Tabellen sehr gebräuchlich, wobei das dahinter stehende Dateiformat selbst eine Tabellen-Datei sein *kann* (z.B.: .csv), aber nicht sein *muss*. Jede Zeile wird dabei mit im Vorhinein definierten Textbestandteilen (Tokens) befüllt und enthält zusätzlich einen eindeutigen Identifikator (z.B. eine fortlaufenden Nummer, beginnend mit 1). Welche Textbestandteile dabei in den einzelnen Zellen und Zeilen der Tabelle stehen (was also als Token gezählt wird) hängt dabei davon ab, was und wie annotiert werden soll. Mit einem elegischen Distichon wie jenem von Matthias Claudius (entnommen aus: <https://de.wikipedia.org/wiki/Distichon>; 09.07.2019):

„Im Hexameter zieht der ästhetische Dudelsack Wind ein.

Im Pentameter drauf lässt er ihn wieder heraus.“

kann man unter anderem wie folgt verfahren:

- (1) Man zählt das Distichon als ganzes als einen Token (seltene Vorgehensweise);
- (2) man zählt jede Zeile als jeweils einen Token (in der Praxis ebenfalls selten);
- (3) man zählt jedes Wort als einzelnen Token, was, wie bereits erwähnt, in der Annotationspraxis häufig geschieht.

Tabellarisch dargestellt ergibt sich hieraus (unter Vernachlässigung von Satzzeichen):

¹²⁸ Hinrichs, Uhle, Linguistische Annotation (wie Anm. 102), 225.

(1)

Identifikator	Token	Metadaten/Annotation (z.B.: Autor/Versmaß)
1	Im Hexameter zieht der ästhetische Dudelsack Wind ein; Im Pentameter drauf lässt er ihn wieder heraus	Matthias Claudius/elegisches Distichon

oder:

(2)

Identifikator	Token	Metadaten/Annotation (z.B.: Versmaß)
1	Im Hexameter zieht der ästhetische Dudelsack Wind ein;	Hexameter
2	Im Pentameter drauf lässt er ihn wieder heraus	Pentameter

oder:

(3)

Identifikator	Token	Metadaten/Annotation (z.B.: Wortarten nach STTS)
1	Im	APPR
2	Hexameter	NN
3	zieht	VVFIN
4	der	ART
5	ästhetische	ADJA
...	...	...

Die tabellarisch-vertikale Tokenisierung eignet sich vor allem „zur Darstellung sequenzieller Informationen und weniger für hierarchische Strukturen.“¹²⁹

Die Beispiele in den jeweils rechten Spalten der Tabellen sollen veranschaulichen, dass Annotation nicht zwangsläufig darin bestehen muss, die Bestandteile eines Textes mit explizierenden Informationen zu ihrer Wortart anzureichern.

¹²⁹ Hinrichs, Uhle, Linguistische Annotation (wie Anm. 102), 226.

Annotation muss, wie eingangs erwähnt, nicht per se linguistischer Natur sein. Vielmehr kann sie, verstanden als übergeordnetes Arbeitsprinzip und als prinzipielle Methode, eine ganze Reihe unterschiedlicher Metadaten am Text generieren.

In Hinblick darauf, dass grundsätzlich ganz unterschiedliche sprachliche Einheiten als Tokens definiert werden können, wird das Problem, wie ein Token gegenüber dem nächsten in dieser sequenziell-tabellarischen Darstellung abzugrenzen ist, hier bewusst nicht mehr näher besprochen. Angemerkt sei lediglich, dass bei der linguistischen Annotation für gewöhnlich als Token definiert wird, was zwischen zwei Leerzeichen steht. Von diesem Prinzip kann aber abgewichen werden, wenn etwa typographische Gewohnheiten in alten Drucken dies erfordern.

Die Darstellung eines Textes als Tabelle kann dabei, wie gesagt, im Dateiformat selbst begründet liegen oder bloß die graphische Benutzeroberfläche eines anderen, „darunter liegenden“ Dateiformates sein.

Zu *dem* bevorzugten „Format“ zur Darstellung digital vorliegender und annotierter Texte schlechthin hat sich innerhalb der Corpuslinguistik, ja innerhalb der Digital Humanities überhaupt in den letzten Jahren vor allem *eine* „Sprache“ entwickelt – die digitale Auszeichnungssprache *XML* (*eXtensible Markup Language*) in spezieller Ausprägung der *TEI* (*Text Encoding Initiative*), also *TEI-XML*.¹³⁰ Der Terminus „Auszeichnung“ ist hier im Sinne von Annotation (also: Informationsexplizierung und/oder Informationsanreicherung mit Metadaten) zu verstehen. (TEI-)XML stellt ein flexibles Regelwerk zur Annotation von Texten dar. Seine Entwicklung wurde 1994 in Abgrenzung und als Erweiterung zur eher statischen Auszeichnungssprache *Hypertext Markup Language* (HTML) initiiert.¹³¹ „The Text Encoding Initiative [...] is a consortium which collectively develops and maintains a standard for the representation of texts in digital form“¹³² – „to represent all kinds of textual material [...]“¹³³

Der Begriff *Standard* ist hier im Sinne einer Sammlung von Empfehlungen zu verstehen. Entwickelt, „um die Strukturen eines Dokuments kenntlich [...] zu

¹³⁰ Vgl. zum Folgenden: Patrick Sahle, Georg Vogeler, XML. In: Jannidis, Kohle, Rehbein, Digital Humanities (wie Anm. 22), 128–146.

¹³¹ Vgl. Rehm, Texttechnologische Grundlagen. In: Carstensen, Ebert et al., Computerlinguistik und Sprachtechnologie (wie Anm. 57 und 21), 161.

¹³² <https://tei-c.org/>; letzter Zugriff: 31.05.2019.

¹³³ <http://www.tei-c.org/About/>; letzter Zugriff: 11.06.2018.

machen, indem [Annotationen direkt] in einen [...] Text eingefügt werden,¹³⁴ ist XML für Menschen wie für den Computer gleichermaßen lesbar, wobei im Gegensatz zum Tabellenformat eine längere Eingewöhnung seitens des menschlichen Bearbeiters/der Bearbeiterin erforderlich ist, da XML-codierte Texte sehr viele Metadaten und Annotationen unterschiedlichster Kategorien enthalten können. Wir haben weiter oben bereits ein Beispiel für einen xml-codierten Text angeführt.

Bei dieser Art von Informationsanreicherung werden sprachliche Einheiten, Tokens, die auch *Strings* (Zeichenfolgen) genannt werden, von korrespondierenden Tag-Paaren „ummantelt“. Das heißt: Der Token „xyz“ wird in der Regel umgeben von einem Anfangs- bzw. einem Öffnungstag und von einem Schluss- bzw. Schließ-Tag. Für Tags in XML ist dabei die Schreibung in spitzen Klammern typisch. Anfangs- und Schließ-Tag müssen außerdem über die gleiche Benennung verfügen, also: <tag>TokenXYZ</tag>. Der Schrägstrich im Schließ-Tag </tag> zeigt an, dass es sich um einen Schließ-Tag handelt. Bezüglich der Schreibweise mit Spitzklammern ist auch von *Mark-up* die Rede. Was innerhalb von Spitzklammern steht, ist Mark-up (im Unterschied zu den Zeichenfolgen zwischen den Tag-Paaren – diese sind, sozusagen, Fließtext). Die Arbeit mit Tag-Paaren bzw. mit Mark-up ist notwendig, um dem Computer den Unterschied zwischen Annotation/Metadaten und Fließtext (sprachliche Primärdaten) zu verdeutlichen.¹³⁵ Die Zeichenketten innerhalb der Spitzklammern müssen mit einem Buchstaben oder mit einem Unterstrich beginnen. Der ummantelte Token und das ihn ummantelnde Tag-Paar bilden zusammen ein *Element*. Ein Element ist nur vollständig mit einem Öffnungs- und einem Schließ-Tag. Jeder einmal geöffnete Tag <x> muss auch wieder mit einem Schließ-Tag </x> geschlossen werden. „Elemente dienen der Abstraktion von Textphänomenen oder der Beschreibung von [Text-]Strukturen.“¹³⁶

Da Tag-Paare direkt an Tokens angebracht, also direkt in den Text „eingebettet“ werden, spricht man in Bezug auf XML generell von einem „*embedded format*“¹³⁷ (im Gegensatz zur *Stand-alone*-Annotation, wo Text und Annotationen in jeweils

¹³⁴ Sahle, Vogeler, XML. In: Jannidis, Kohle, Rehbein, Digital Humanities (wie Anm. 21), 128.

¹³⁵ Vgl. Fotis Jannidis, Grundlagen der Datenmodellierung. In: Jannidis, Kohle, Rehbein, Digital Humanities (wie Anm. 21), 99–108; hier: 99.

¹³⁶ Sahle, Vogeler, XML. In: Jannidis, Kohle, Rehbein, Digital Humanities (wie Anm. 21), 135.

¹³⁷ Jensen, McGillivray, Quant. Hist. Ling. (wie Anm. 58), 108.

separaten Dateien gespeichert werden)¹³⁸. Elemente können Tokens und/oder weitere Elemente enthalten, so etwa:

```
<name type="person"><vorname>Leopold</vorname>  
<nachname>Mozart</nachname></name>.
```

Zusätzlich zur hier gewählten, dem Textfluss dieser Arbeit geschuldeten sequenziellen Darstellungsweise der Tags und der Elemente stehen diese auch in einem strikt hierarchischen Verhältnis zueinander. Sie können daher in Form eines Baumes dargestellt werden, das heißt: Nicht jedes beliebige Element darf innerhalb jedes beliebigen Elementes stehen. Das Tag-Paar `<name>...</name>` muss den Tag-Paaren für Vor- und für Nachname übergeordnet sein.

Innerhalb des Schemas, das die TEI für XML vorschlägt, folgt die Hierarchie der Tags untereinander teilweise der Logik und der Struktur der „textuell-sprachlichen Wirklichkeit“, die in XML abgebildet werden soll: Die Tags für das Phänomen „Strophe“ sind den Tags für das Phänomen „Gedicht“ untergeordnet. Teilweise gehorcht diese Hierarchie aber auch solchen Prinzipien, die dem TEI-Standard selbst eigen sind. Die Anordnung der Tags muss (unabhängig vom konkreten Einzelstandard) in jedem Fall entsprechend den Prinzipien der Unter- und der Überordnung vorgenommen werden. Man sagt, dass die Einbettung bzw. die Verschachtelung (das „*nesting*“) der Tags korrekt sein muss.

Hinsichtlich Tags, die einander über- und untergeordnet sind, spricht man bei direkter hierarchischer Abhängigkeit vom Eltern-Kind-Verhältnis, bei Tags, die in vertikaler Ebene weiter voneinander entfernt sind, von Vorfahre und Nachkomme. Tags, die innerhalb der Hierarchie auf derselben Stufe stehen und dasselbe übergeordnete Elternelement aufweisen, nennt man Geschwister-Tags.

Auf diese Weise kann in XML jede erdenkliche Information, die einem Text hinzugefügt werden soll oder die in diesem bereits implizit enthalten ist, (vor allem für den Computer) explizit lesbar gemacht werden (für den menschlichen Bearbeiter/die Bearbeiterin ist vieles auch ohne Annotation bereits explizit kenntlich), etwa, dass es sich bei einer bestimmten Zeichenfolge um einen

¹³⁸ *Jenset, McGillivray*, Quant. Hist. Ling. (wie Anm. 58), 110.

Tagebucheintrag handelt, dass eine bestimmte Zeile eine Datumszeile oder eine Überschrift darstellt, dass ein bestimmter Text als `<gedicht>...</gedicht>` zu identifizieren ist und innerhalb des Textes bestimmte Zeichengruppen als Strophen zu verstehen sind, um nur ein paar Beispiele zu nennen.

All dies lässt sich in XML aber nicht nur mit Elementen allein bewerkstelligen, sondern, wie im obigen Beispiel bereits angeführt, auch mit sogenannten *Attributen* innerhalb eines Tags. Attribute spezifizieren näherhin, wofür ein Tag steht. Im Element `<name type=„person“> ... </name>` spezifiziert das Attribut `type=„person“` die Zeichenfolge bzw. die Entität zwischen den Tag-Paaren (oben: Leopold Mozart) näher: Es kennzeichnet sie in diesem Fall als Person.

Unter dem Begriff Entität versteht man in der Informatik ein eindeutig identifizierbares Objekt bzw. einen eindeutig identifizierbaren Sachverhalt (materiell oder immateriell), über das bzw. über den Informationen gespeichert oder verarbeitet werden sollen. Entitäten, die in den Digital Humanities mit XML häufig als solche kenntlich gemacht werden, sind beispielsweise: Personen, Orte, Institutionen, Werktitel, wobei die Definition, was als Entität gilt und als solche annotiert wird, auch vom jeweiligen Bearbeiter/der Bearbeiterin des Textes abhängig sein kann. Attribute sind grundsätzlich Teil der Öffnungstags und bestehen aus einem *Attributnamen* (vor dem Gleichheitszeichen; hier: `type`) und einem *Attributwert* (in Anführungszeichen nach dem Gleichheitszeichen; hier: `„person“`), ohne, dass zwischen Attributname und Attributwert durch Leerzeichen getrennt würde. Innerhalb eines Öffnungstags kann es eine ganze Reihe von Attributen geben, sofern diese durch Leerzeichen voneinander abgegrenzt sind und nicht denselben Attributnamen aufweisen.

Ein großer Vorteil von XML besteht darin, dass es erlaubt ist, eigene Tags bzw. Elemente je nach Bedarf zu kreieren, je nach dem, was und wie annotiert werden soll. Im Regelfall definiert man „Elemente für Phänomene, die [in einem Text] voraussichtlich häufiger vorkommen bzw. auf mehrere Texte oder Dokumente anwendbar sind.“¹³⁹ Die Zeichenfolgen zwischen den Spitzklammern, also die Namen der Tags, sollten dabei möglichst selbsterklärend und allgemein verständlich sein, sodass deutlich wird, wofür ein Tag steht.

¹³⁹ Sahle, Vogeler, XML. In: Jannidis, Kohle, Rehbein, Digital Humanities (wie Anm. 21), 135.

Dank seiner Flexibilität und der nahezu unendlichen Möglichkeit der individuellen Erweiterung bietet XML eine ausgezeichnete Ausgangsbasis für Annotationen und/oder für editorisches Arbeiten im digitalen Umfeld, zumal es den Erfordernissen unterschiedlichster editorischer Unternehmungen angepasst werden kann.

2) Praktischer Hauptteil

Der Hauptteil vorliegender Arbeit verfolgt zwei Zielsetzungen. Einerseits soll ein evaluierender Überblick geschaffen werden über jene Tagsets, die, soweit dem Verfasser dieser Zeilen bekannt, gegenwärtig für die linguistische Annotation lateinischer Texte zur Verfügung stehen (Stand Juni 2019).

Zweitens werden darin, wie eingangs angekündigt, jene Erfahrungen verschriftlicht, die sich im Rahmen der praktischen Annotation der Briefe Eckharts an die Brüder Pez ergaben.

2.1) Lateinische Tagsets für PoS und Morphologie

2.1.1) Allgemeine Vorbemerkungen

Soweit dem Verfasser vorliegender Arbeit bekannt, steht für morphologische, syntaktische und für PoS-Annotation lateinischer Texte gegenwärtig (Stand: Juni 2019) eine Reihe von Tagsets zur Verfügung. Nicht alle können im selben Maße als eigenständig bezeichnet werden. Sie wurden im Rahmen wissenschaftlicher Projekte jeweils unterschiedlicher Dimension erprobt und zumeist auch anhand von schriftlichen Publikationen dokumentiert. Letzteres – der Aspekt der schriftlichen Dokumentation – erweist sich jedoch nicht allen Fällen als befriedigend.

Ihrer Funktion nach können diese Tagsets grob in zwei Gruppen eingeteilt werden: Einerseits handelt es sich um solche, die allein der morphologischen und/oder der PoS-Annotation dienen, andererseits um solche, die zur Erstellung von Treebanks (Baumbanken) vorgesehen sind, also zur syntaktischen Annotation geparster Texte.¹⁴⁰ Erstere werden in Kürze vorgestellt. Einen „De-facto-

¹⁴⁰ Hinsichtlich der Anzahl ist nicht auszuschließen, dass im Rahmen kleinerer Projekte bereits weitere Tagsets für die lateinische Sprache entwickelt wurden, die dem Verfasser der

Standard“, also *ein* lateinisches Tagset, das sich auf Grund seiner Art und auf Grund seiner Verwendung im Rahmen mehrerer wissenschaftlicher Projekte gleichsam als kanonisch für *alle* Annotationsarten durchgesetzt hätte, wie das im Falle des STTS zum Zweck der PoS-Annotation deutscher Texte zu beobachten ist, gibt es dabei nicht. Dies liegt freilich auch daran, dass Tagsets für PoS- und für morphologische Annotation gänzlich anderen Ansprüchen Genüge tun müssen verglichen mit solchen, die der syntaktischen Annotation dienen.

Beide Arten von Tagsets sind wegen unterschiedlicher Zielsetzungen nur sehr begrenzt miteinander vergleichbar. Umgekehrt ist zu betonen, dass die großen „Flugschiffprojekte“ der DH, die in den letzten Jahren und Jahrzehnten mit lateinischen Texten befasst waren (Index Thomisticus Treebank, Latin Dependency Treebank und PROIEL), allesamt sowohl PoS- wie auch morphologisch *und* syntaktisch annotierte Texte bieten (in jeweils unterschiedlicher Quantität und Qualität) – syntaktisch gekennzeichnete Texte eben in Form von Treebanks. Zusätzlich sind die angesprochenen Projekte über die Jahre beträchtlich gewachsen, äußerst vielschichtig geworden und haben fallweise eigene Seitenstränge und Teilprojekte kleinerer Dimension entwickelt. Folglich fällt die Unterscheidung, welches Tagset innerhalb eines Großprojektes für welche verschiedenen Teilaufgaben verwendet wurde trotz der Unterschiedlichkeit der Teilaufgaben auf den ersten Blick nicht immer leicht. Die Anzahl der Publikationen, die alleine im Rahmen jener drei genannten Flugschiffprojekte verfasst wurden, ist, wie weiter unten noch zu zeigen sein wird, enorm groß.

2.1.2) Tagging-Tools, Tagsets und Modelle

Hinsichtlich automatisierter Annotation ist die Einsetzbarkeit der etablierten Tagsets zumeist an die Verwendung bestimmter digitaler Tools (Tagger oder Parser) gebunden, die speziell für den Einsatz eines oder mehrerer jener Tagsets trainiert und/oder sogar dafür entwickelt wurden. Tagger sind, wie gesagt, für die PoS- und für die morphologische Annotation beliebiger Sprachen zu verwenden, Parser für die syntaktische. Viele dieser Tools stehen im Internet zwar öffentlich

vorliegenden Zeilen gegenwärtig jedoch nicht greifbar sind.

und kostenlos zur Verfügung (als Download¹⁴¹ oder als Web-Application über den Browser), doch sind die Möglichkeiten, Tagsets und Tools frei miteinander zu kombinieren aus eben den genannten (technischen) Gründen beschränkt und an Voraussetzungen gebunden: Denn bevor automatisch annotiert werden kann, muss ein Tool, wie bereits angedeutet, oftmals an ein Tagset gewöhnt (darauf trainiert) werden – und das mittels eines manuell vorannotierten Textes, eines Trainingscorpus,¹⁴² dessen Erstellung äußerst zeitaufwendig ausfallen kann, je nach Umfang des Textes und je nach Tagset, je nach Annotationslevel und Annotationstiefe.

Für die Erstellung des Trainingscorpus wird einerseits ein Text, andererseits ein Tagset ausgewählt bzw. aufgesetzt. Anhand der damit (manuell!) vorannotierten Texte erstellt der Tagger ein *Modell*. Der Tagger versucht also, grob gesprochen, jene Regeln zu analysieren und zu „durchschauen“, mit denen das Trainingscorpus manuell vorannotiert wurde, um diese Regeln in der Folge auf weitere, ungetaggte Texte anzuwenden. Das Annotationsmodell dient somit als Vorlage. Die Qualität automatischer Annotation hängt demnach, wiederum vergrößert gesprochen, von mehreren Faktoren ab: einerseits von jener des ausgewählten Tagsets selbst, andererseits von der Richtigkeit der manuellen Annotation im Trainingscorpus, drittens von den Kapazitäten und der Trefferquote des Taggers, woraus das Modell für die automatische Annotation entsteht, das – viertens – allenfalls nachkorrigiert werden kann. Tagger jüngerer Generation arbeiten, wie weiter oben bereits angedeutet, vielfach auf Basis von statistischen Verfahren. Auch sie müssen in der beschriebenen Weise trainiert werden.

In Form des Modells entsteht während der Trainingsphase eines Taggers auf diesem Wege also eine enge Verbindung zwischen Tagset und Tool. Diese enge Verbindung zwischen Tagset und Tool, die auf diese Weise zu Stande kommt, macht die daraus entstehende Kombination – das Modell – für den User/die Userin gleichsam zu einem proprietären Gut, ungeachtet der öffentlichen Verfügbarkeit beliebiger Tagsets und beliebiger Tools über das Internet. In der

¹⁴¹ Vgl. e.g. <https://nlp.stanford.edu/links/statnlp.html>; letzter Zugriff: 30.05.2019.

¹⁴² Vgl. e.g. *Jenset, McGillivray*, Quant. Hist. Ling. (wie Anm. 58), 113.

Praxis – während der Anwendungsphase – wird man sich daher vielfach mit Kompromissen behelfen, denn nicht immer sind die besten Tagger, die die höchsten automatischen Trefferquoten erzielen könnten, auf jene Tagsets abgestimmt, das aus latinistischer Sicht die besten wären.

Interoperabilität zwischen unterschiedlichen Tagsets und Tools ist grundsätzlich dann gegeben, wenn mehrere digitale Werkzeuge mit Tag-Inventaren arbeiten, die auf dieselben linguistischen Kategorien rekurren, wobei diese je nach Tagset unterschiedlich benannt sein können, während sie gleichzeitig für ein und dasselbe linguistische Phänomen stehen. Die Benennungen der Tags sind in diesem Fall sekundär. Im Rahmen der vorliegenden Arbeit wurden keine Versuche zur Neukombination von Tagging-Tools und Tagsets durchgeführt, da dies Arbeiten technischer Natur erfordern würde, die im Rahmen dieser Untersuchung nicht möglich sind. Vielmehr wurde auf bestehende Lösungen zurückgegriffen. Im Falle einer Neukombination empfiehlt es sich, eine Kooperation mit entsprechenden Fachleuten anzustreben.

Als die besten Tagger für die PoS- und für die morphologische Annotation lateinischer Texte gelten derzeit unter anderem MarMot, FLORS¹⁴³ oder Lapos¹⁴⁴. Im Rahmen unterschiedlicher Aufgabenstellungen (Lemmatisierung, automatisiertes PoS-Tagging, automatische Annotation von morphologischen Kennzeichen) wurden diese Tools an lateinischen Textcorpora (hauptsächlich des Mittelalters) in den letzten Jahren mehrfach trainiert und getestet.¹⁴⁵ Hierbei wird im Allgemeinen

¹⁴³ Tobias Schnabel, Hinrich Schütze, FLORS: Fast and Simple Domain Adaption for Part-of-Speech Tagging. In: *Association for Computational Linguistics*, Sharon Goldwater (Hgg.), Transactions of the Association for Computational Linguistics 2 (2014), 15–26; online unter: <http://acl2014.org/acl2014/Q14/pdf/Q14-1005.pdf>; letzter Zugriff: 31.05.2019.

¹⁴⁴ Vgl. Yoshimasa Tsuruoka, Yusuke Miyao, Jun'ichi Kazama, Learning with Lookahead. Can History-Based Models Rival Globally Optimized Models? In: *Association for Computational Linguistics* (Hg.), Proceedings of the Fifteenth Conference on Computational Natural Language Learning (Portland 2011), 238–246; online unter: <http://www.aclweb.org/anthology/W11-0328>; letzter Zugriff: 31.05.2019.

¹⁴⁵ Zur Performance dieser Tagger vgl. insbesondere Steffen Eger, Rüdiger Gleim, Alexander Mehler, Lemmatization and Morphological Tagging in German and Latin. A Comparison and a Survey of the State-of-the-Art. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (Portorož 2016), 1507–1513; online unter: <https://pdfs.semanticscholar.org/99b3/691ac04d7a196be1b7115189aca51f143c6d.pdf>; letzter Zugriff: 31.05.2019;
zur Performance von Lapos vgl. neben dem soeben zitierten Artikel auch: Steffen Eger, Alexander Mehler, Tim von der Brück, Lexicon-Assisted Tagging and Lemmatization in Latin: A Comparison of six Taggers and two Lemmatization Methods. In: *Association for Computational Linguistics, The Asian Federation of Natural Language Processing* (Hgg.), Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (Beijing 2015), 105–113; online unter:

für die Verwendung neuerer Tools plädiert, entgegen der weit verbreiteten Verwendung des älteren TreeTaggers, dem eine gewisse Pionier-Rolle zuzukommen scheint: „[M]ore recent tagger classes substantially outperform their predecessor generation.“¹⁴⁶

Allerdings wurden diese Tests teilweise mit stark reduzierten Tagsets vorgenommen, deren Bestandteile bei weitem nicht jede Wortart exakt abdecken, die im Lateinischen vorkommt. Auch hier hat man es immer wieder mit Kompromissen zu tun.¹⁴⁷ Neben vielen slawischen Sprachen, dem Altgriechischen und dem Deutschen zeichnet Latein sich nämlich durch beträchtlichen morphologischen Reichtum aus (viele unterschiedliche Endungen!). Dies stellt im Bereich des Natural Language Processing neben der relativ freien Wortstellung ein bekanntes Problem dar¹⁴⁸, erstens für das Training der automatischen Annotationstools selbst, zweitens für das Design der Tagsets insofern, als diese nicht zu lang werden sollten, damit die Zeit, die für das Training der Tools benötigt wird, dem Umfang nach vertretbar bleibt.

Mit dem STTS liegt im Falle des Deutschen jedoch ein Tagset für eine morphologisch reiche Sprache vor, das zwar nicht als vollständig¹⁴⁹, wohl aber als hochdifferenziert bezeichnet werden darf und das neben Informationen zur Wortart vereinzelt auch morphologische Merkmale expliziert.

Dass also gerade für das Lateinische mit seinem morphologischen Reichtum Tagsets verwendet werden, die fallweise nicht sonderlich differenziert erscheinen,

<http://www.aclweb.org/anthology/W15-3716>; letzter Zugriff: 31.05.2019; über die Performance weiterer Tools berichtet *McGillivray*, *Methods in Latin Computational Linguistics* (wie Anm. 50), 22 ff.

¹⁴⁶ *Eger, Mehler, vor der Brück*, *Lexicon-Assisted Tagging and Lemmatization* (wie Anm. 145), 105.

¹⁴⁷ Das von *Eger, Mehler, vor der Brück*, *Lexicon-assisted tagging and lemmatization* (wie Anm. 145), 106–109 verwendete „Frankfurter-Tagset“, auf das noch näher einzugehen sein wird, kennt nur eine einzige Art Pronomina und unterscheidet nicht zwischen bei- und unterordnenden Konjunktionen. Gleichwohl heißt es ebd., 106 in Bezug auf das im Hintergrund der Annotation verwendete Lexikon: „Pronouns are annotated with a pronoun type that further differentiates pronouns into demonstrative, interrogative, personal, reflexive, relative, possessive, indefinite, intense, and correlative pronouns.“ Die dafür vorgesehenen Tags sind der zitierten Publikation allerdings nicht zu entnehmen oder nicht in dieser Genauigkeit in die Tagging-Tests eingeflossen.

¹⁴⁸ Vgl. *Eger, Mehler, vor der Brück*, *Lexicon-Assisted Tagging & Lemmatization* (Anm. 145), 105.

¹⁴⁹ Im STTS fehlt beispielsweise ein eigener Tag für das Präsenspartizip, für hauptwörtlich gebrauchte Adjektive oder Infinitive.

ist vor diesem Hintergrund nur teilweise nachzuvollziehen, zumal die einschlägige Forschung auch das Tagging morphologisch reicher Sprachen zusehends in den Griff bekommt.¹⁵⁰

Insgesamt ergibt sich demnach ein höchst zwiespältiges Bild: Hinsichtlich gegenwärtig verfügbarer Tagsets für PoS- und für morphologisches Tagging lateinischer Texte ist die Spanne zwischen differenzierten und reduzierten Varianten extrem breit. Gleichwohl existiert in der automatisierten Annotationspraxis die Möglichkeit, technischen Problemen, die durch die Verwendung stark ausdifferenzierter Tagsets entstehen, insofern zu begegnen, als man einen „multi-stage tagging approach [wählen kann,] in which tagging is initially [!] performed with a reduced tagset.“¹⁵¹ Die einzelnen Schritte zur vollständigen linguistischen Annotation (Wortarten einerseits und morphologische Merkmale andererseits) werden fallweise getrennt und separat durchgeführt, wobei die Tagsets offenbar je nach Bedarf und je nach Fortschritt des Annotationsprozesses erweitert werden können: „Contrary to some of our related work, we view the morphological tagging problem for Latin as a multi-label tagging problem in which each tagging task (PoS, case, gender, et c.) is handled independently.“¹⁵²

Gleichwohl erscheinen manche Tagsets für das Latein nicht einmal würdig, als Teil einer Mittelschulgrammatik zu fungieren. Schlecht ist es zuweilen auch um die Auffindbarkeit der Tagsets in der Literatur bzw. im Internet bestellt, schlecht auch um die Nachvollziehbarkeit, wer im Endeffekt für die Erstellung eines Tagsets verantwortlich zeichnet. Denn innerhalb der Computational Humanities, der Computerlinguistik sowie innerhalb der Texttechnologie liegt der Schwerpunkt des Interesses naturgemäß stark im Bereich der Annotationstools, also im Bereich der Tagger und der Parser selbst, sowie im Bereich ihres Designs, ihres Verhaltens, ihrer Resultate bei verschiedenen Sprachen. Und während die Auszeichnung von Wortarten und von morphologischen Kennzeichen im Rahmen der digitalen Editionsphilologien an Bedeutung gewinnt und weiter gewinnen kann, stehen bzw. standen PoS- und morphologische Annotation innerhalb der computerassoziierten Sprachwissenschaften häufig im Dienste (bzw. im Schatten)

¹⁵⁰ Vgl. Müller, Schmid, Schütze, Efficient Higher-Order CRFs (wie Anm. 16).

¹⁵¹ Eger, Mehler, vor der Brück, Lexicon-Assisted Tagging and Lemmatization (wie Anm. 145), 106.

¹⁵² Eger, Mehler, vor der Brück, Lexicon-Assisted Tagging and Lemmatization (wie Anm. 145), 108.

von Projekten, die sich primär dem Treebanking widme(te)n, also der Erstellung von syntaktisch annotierten Corpora. Bei solchen Projekten stehen naturgemäß jene Tagsets im Vordergrund, die der syntaktischen Annotation dienen, was, wie gesagt, insgesamt dazu führt, dass Tagsets für Wortarten- und für morphologische Annotation teils schwierig zu finden sind.

Und es tritt noch ein weiterer Faktor erschwerend hinzu: Innerhalb jener größeren und äußerst repräsentativen Projekte, die sich gleichsam am Schnittpunkt zwischen lateinischer DH-Philologie und Computerlinguistik befinden, also beispielsweise innerhalb der Perseus Digital Library einschließlich ihrer assoziierten Projekte (Latin Dependency Treebank) sowie im Rahmen des Index Thomisticus, hat man vielfach zusammengearbeitet¹⁵³, sich verständlicherweise mannigfacher Synergien bedient und sich aufeinander berufen. Dies führt, wie bereits angedeutet, dazu, dass die eigentlichen Urheber mancher lateinischer Tagsets nur mehr schwer zu eruieren bzw. nur mehr schwer von einander zu unterscheiden sind.

¹⁵³ Vgl. David *Bamman*, Gregory *Crane*, Marco *Passarotti*, Savina *Raynaud*, A Collaborative Model of Treebank Development. In: Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (Bergen 2007), 1–6; online unter: https://www.researchgate.net/publication/28584823_A_Collaborative_Model_of_Treebank_Development; letzter Zugriff: 31.05.2019; vgl. ebenso: David *Bamman*, Roberto *Busa*, Gregory *Crane*, Marco *Passarotti*, The Annotation Guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The Treatment of some Specific Syntactic Constructions in Latin. In: Proceedings of the LREC 2008 (Marrakech 2008), 71–76; online unter: http://www.lrec-conf.org/proceedings/lrec2008/pdf/25_paper.pdf; letzter Zugriff: 31.05.2019.

2.1.2.1) Das Morpheus Morphological Tagset

1: part of speech

n	noun
v	verb
t	participle
a	adjective
d	adverb
c	conjunction
r	preposition
p	pronoun
m	numeral
i	interjection
e	exclamation

2: person

1	first person
2	second person
3	third person

3: number

s	singular
p	plural

4: tense

p	present
i	imperfect
r	perfect
l	pluperfect
t	future perfect
f	future

5: mood

i	indicative
s	subjunctive
n	infinitive
m	imperative

p participle
g gerundive¹⁵⁴
u supine

6: voice

a active
p passive

7: gender

m masculine
f feminine
n neuter

8: case

n nominative
g genitive
d dative
a accusative
b ablative
v vocative
l locative
i instrumental

9: degree

c comparative
s superlative

e.g.: alium: a-s---ma-

1: a adjective
2: -
3: s singular
4: -
5: -

¹⁵⁴ Ein eigener Tag für das Gerundium wurde, wie online ersichtlich, mittlerweile ergänzt, zumindest im Falle der mit der Perseus Library assoziierten Treebanks; vgl. https://perseusdl.github.io/treebank_data/; letzter Zugriff: 31.05.2019.

- 6: -
- 7: m masculine
- 8: a accusative
- 9: -

Dieses Tagset¹⁵⁵ wurde im Rahmen der Perseus Digital Library¹⁵⁶ sowie im Rahmen der Latin Dependency Treebank¹⁵⁷ entwickelt und dort für das Wortarten-Tagging sowie für die morphologische Annotation verwendet. Hinsichtlich der Latin Dependency Treebank ist es ausdrücklich abzugrenzen von jenem Tagset, das zum Einsatz kam, um Baumbanken zu erstellen, also für die syntaktische Annotation.¹⁵⁸ Diesem syntaktischen Tagset liegt seinerseits das System der Prague Dependency Treebank zu Grunde,¹⁵⁹ das nicht näherhin Gegenstand der vorliegenden Untersuchung sein soll. Auch die lateinischen Texte, die innerhalb

¹⁵⁵ Online unter: https://itreebank.marginalia.it/doc/Tagset_Perseus.pdf;
letzter Zugriff: 31.05.2019.

¹⁵⁶ Vgl. Gregory *Crane*, Building a Digital Library. The Perseus Project as a Case Study in the Humanities. In: Proceedings of the 1st ACM International Conference on Digital Libraries (New York 1996), 3–10; online unter: https://www.researchgate.net/publication/234791911_Building_a_digital_library_The_Perseus_Project_as_a_case_study_in_the_humanities;
vgl. ebenso: Gregory *Crane*, Cultural Heritage Digital Libraries. Needs and Components. In: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (London 2002), 626–663; online unter: <http://www.perseus.tufts.edu/~ababeu/ecdl2002.pdf>;
ebenso: Gregory *Crane*, Clifford *Wulfsberg* et al., Towards a Cultural Heritage Digital Library. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) (Washington 2003), 75–86; online unter: <http://www.ccs.neu.edu/home/dasmith/jcdl2003.pdf>;
letzter Zugriff auf alle URLs: 31.05. 2019.

¹⁵⁷ Vgl. David *Bamman*, Gregory *Crane*, The Design and Use of Latin Dependency Treebank. In: Jan *Hajič*, Joakim *Nivre* (Hgg.), Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (Prag 2006), 67–78; Artikel online unter: <http://ufal.mff.cuni.cz/tlt2006/pdf/110.pdf>;
Webpräsenz des Projektes: https://perseusdl.github.io/treebank_data/;
vgl. ebenso: David *Bamman*, Gregory *Crane*, The Latin Dependency Treebank in a Cultural Heritage Digital Library. In: *Association for Computational Linguistics* (Hg.), Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data – LaTeCH 2007 (Prag 2007), 33–40; online unter: <https://www.aclweb.org/anthology/W07-0905>; und:
David *Bamman*, Gregory *Crane*, Structured Knowledge for Low-Resource Languages: The Latin and Greek Dependency Treebanks. In: Proceedings of the Text Mining Services 2009 (Leipzig 2009), 1–10; online unter: <https://pdfs.semanticscholar.org/2b86/9b06c81b19c2a76bcdcbd7dba5a6aad2f672.pdf>;
letzter Zugriff auf alle URLs: 31.05.2019.

¹⁵⁸ David *Bamman*, Gregory *Crane*, Marco *Passarotti*, Savina *Raynaud*, Guidelines for the Syntactic Annotation of Latin Treebanks, Version 1.3 (Tufts/Medford 2007); online unter: https://itreebank.marginalia.it/doc/2007_Passa+Bamman+Crane+Raynaud_Guidelines%20Tb.pdf; letzter Zugriff: 31.05.2019.

¹⁵⁹ Vgl. Jan *Hajič*, Building a Syntactically Annotated Corpus. The Prague Dependency Treebank. In: Eva *Hajičová* (Hg.), Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová (Prag 1998), 12–19; Web-Präsenz des Projektes: <https://ufal.mff.cuni.cz/pdt3.5>; letzter Zugriff auf beide URLs: 31.05.2019; vgl. auch Anm. 188.

der Perseus Digital Library verfügbar sind, wurden, wie gesagt, mit dem Morpheus-Tagset annotiert.¹⁶⁰ Es deckt einen großen Teil der lateinischen Wortarten ab und enthält Tags für fast alle morphologischen Merkmale.

Nicht unterschieden wird zwischen beiordnender und unterordnender Konjunktion. Eine gezielte Suche nach Haupt- bzw. nach Nebensätzen ist somit nur innerhalb jener Perseus-Texte möglich, zu denen es auch Baumbanken gibt. Eigene Tags für *named entities* (also für Eigennamen, z.B. von Personen oder von Orten) sind ebenso wenig enthalten. Das stellt aus editionswissenschaftlicher Sicht einen gravierenden Mangel dar. Deklinations- und Konjugationsklassen werden nicht als solche benannt. Für die verschiedenen Arten von Pronomina steht nur ein Tag (1p) zur Verfügung. Man könnte exakter unterscheiden zwischen Supinum I, Supinum II, Imperativ I und Imperativ II (der gleichwohl selten vorkommt). Es gibt keinen Tag für fremdsprachige Tokens. Fremdworte müssen demnach konsequent ihrer Wortart nach annotiert werden, was einen gangbaren Weg bei der Annotation darstellt. Allerdings lassen Fremdworte sich dadurch nicht explizit als solche anzeigen, wenn systematisch nach ihnen gesucht wird.

Eigene Tags für die Verbklasse der Deponentia sowie für die Semideponentia fehlen ebenso. Zwischen Grundzahlworten, Ordnungszahlworten, Distributiva – singuli, bini, terni, also: je ein/-e, je zwei et c. – wird nicht unterschieden. Postpositionen werden nicht als solche ausgewiesen. Auffällig erscheint, dass die Kategorie *participle* zweimal vertreten ist (in 1t und in 5p), die Kategorie Enklitikon (für Partikel wie -ve, -ne, -ce, -que) jedoch gar nicht. Letztere wird in der Latin Dependency Treebank jedoch an die syntaktische Annotation ausgelagert.

Insgesamt stellt das Morpheus Tagset einen Kompromiss zu Gunsten der Übersichtlichkeit und der leichten maschinellen Handhabbarkeit dar, einen Kompromiss, der fallweise etwas minimalistisch wirkt. Immerhin wird der Versuch unternommen, Wortarten- und morphologische Annotation miteinander zu vereinen. Sprachwissenschaftlich gesehen kommt man mit den vorhandenen Tags vorbehaltlich gewisser Einschränkungen im Wesentlichen aus.

¹⁶⁰ Vgl. *Bamman, Crane*, Dynamic Lexicon (wie Anm. 104); Webpräsenz des Projektes: <http://www.perseus.tufts.edu/hopper/> bzw. mit neuem User-Interface seit Frühjahr 2018: <https://scaife.perseus.org/>; letzter Zugriff auf beide URLs: 31.05.2019.

Aus Sicht der Editionswissenschaft ist das Fehlen eines eigenen Tags für Eigennamen am schwersten zu bemängeln. Sie müssten ergänzt werden.

Ein großer Nachteil des Morpheus Tagsets besteht ferner darin, dass die Tags in Kleinbuchstaben und in Ziffern kodiert sind, die je nach Position (!), derer es neun mögliche gibt, eine andere Bedeutung annehmen können (vgl. obiges Beispiel mit *alium*): Die Sigle *a* steht in der Position 1 für Adjektiv, in der Position 8 für Akkusativ. Für sich genommen sind die verwendeten Tags also weder unzweideutig noch selbsterklärend. Diesem Problem begegnete man im Rahmen der Perseus Digital Library und im Rahmen der Latin Dependency Treebank dadurch, dass man die Tags auf der Ebene der Benutzeroberfläche in Form eingängigerer Abkürzungen explizierte (aus „*transeant: v3ppsa---*“ wird dadurch: „*transeant: verb. 3rd.pl.pr.sub.act*“). Dies macht bei der computerbasierten Prozessierung der annotierten sprachlichen Primärdaten zumindest einen zusätzlichen Schritt erforderlich, dessen Aufwand aber nicht besonders groß ist und der damit durchaus vertretbar erscheint. Als Annotationstool der Latin Dependency Treebank fungierte zumeist Arethusa. Dabei handelt es sich um eine frei verfügbare Web-Applikation, die auf Angular JS-javascript basiert – „a back-end independent plugin infrastructure“¹⁶¹.

2.1.2.2) Das Tagset Lamap

Parte del discorso (Wortart/PoS)	tag	Beispiel
Verbo ausiliare “Essere”: Indicativo	ESSE:IND	est
Verbo ausiliare “Essere”: Congiuntivo	ESSE:SUB	sit
Verbo ausiliare “Essere”: Infinito	ESSE:INF	esse
Verbo: Indicativo	V:IND	prohibetur
Verbo: Congiuntivo	V:SUB	addicat
Verbo: Infinito	V:INF	sapere
Verbo: Gerundio	V:GER	lucendum
Verbo: Gerundivo	V:GED	reddendam
Verbo: Participio	V:PTC: nom	sectantes

¹⁶¹ Web-Präsenz abrufbar unter: <http://www.perseids.org/tools/arethusa/app/>; für die Annotation eigener Texte mit Arethusa: <http://www.perseids.org/apps/treebank/>; die auf diesem Wege gewonnenen Daten werden auf dem entsprechenden Server gespeichert und sind abrufbar unter: <http://sosol.perseids.org/sosol/signin>; letzter Zugriff auf alle URLs: 31.05.2019.

	V:PTC: acc	persuasum
	V:PTC: abl	instituto
	V:PTC	perfecte
Verbo: Supino	V:SUP: acc	petitum
	V:SUP:abl	auditu
Verbo: Imperativo	V:IMP	Exaudi
Pronomi	PRON	te
Pronomi: Relativi	REL	quod
Possessivi (Pronomi e/o Aggettivi)	POSS	vestros
Dimostrativi (Pronomi e/o Aggettivi)	DIMOS	Hae
Pronomi: Indefiniti	INDEF	aliquis
Nomi	N:nom	salus
	N:dat	nomini
	N:gen	sitis
	N:loc	domi
	N:acc	calicem
	N:abl	silentio
	N:voc	Christe
Numerali (tutti i tipi)	ADJ:NUM	3
Congiunzioni Coordinanti	CC	et
Congiunzioni Subordinanti	CS	ne
Nomi Propri	NPR	Eph [<i>sic!</i>]
Aggettivi	ADJ	Gregorii [<i>sic!</i>]
	ADJ:COM	minor
	ADJ:SUP	carissimi
	ADJ:abl	toto
Avverbi	ADV	forsitan
Preposizioni	PREP	ab
Interiezioni	INT	Vae
Abbreviazioni	ABBR	R/ [<i>sic!</i>]
Esclamazioni	EXCL	Ma
Parole straniere	FW	effatha
Punteggiatura di fine frase	SENT	.

Punteggiatura NON di fine frase	PUN	,
Simboli	SYM	8. [sic!]
Enclitiche	CLI	que

Das Tagset Lamap¹⁶² wurde von Gabriele Brandolini zur Annotierung lateinischer Texte mit dem Annotationstool TreeTagger entwickelt, für dessen Design Helmut Schmid hauptverantwortlich zeichnet.¹⁶³ Für die Erstellung des Trainingscorpus wurden Texte der klassischen Latinität, der Vulgata und solche aus den Werken des Thomas von Aquin herangezogen.¹⁶⁴ Der TreeTagger dient neben der Erstellung von Lemmata (Grundworten > Lemmatisierung) in erster Linie zur Annotation von Wortarten, wofür folglich auch Lamap entsprechend konzipiert wurde. Andererseits sind darin einige Tags enthalten, die gleichzeitig morphologische Detailinformationen explizieren, beispielsweise Imperativ, Indikativ oder Konjunktiv. Es werden jedoch bei Weitem nicht alle morphologischen Phänomene abgedeckt.

Weder bei Substantiven noch bei Adjektiven noch bei Partizipien oder Pronomina lässt sich deren Numerus kennzeichnen. Tags für Tempora fehlen gänzlich. Und solche für Casus sind zwar vorhanden, hauptsächlich bei Substantiven und bei Partizipien, allerdings äußerst lückenhaft, bei Partizipien besonders unvollständig – es fehlen in diesem Fall zumindest Genitiv und Dativ. Die Funktion des Tags V:PTC mit dem Beispiel *perfecte* ist in diesem Zusammenhang unklar (Partizip? Vokativ?). Auch Pronomina und Adjektive können bekanntlich vollständig im Singular und im Plural dekliniert werden, Letztere in allen Graden (Grundstufe, Komparativ und Superlativ). Im Falle dieser beiden Wortarten stehen jedoch keine eigenen Tags für die Annotation der Casus zur Verfügung – einzig ADJ:abl, vermutlich zur Kennzeichnung von Adjektiven innerhalb eines Ablativus absolutus. Bei Adverbien fehlt die Möglichkeit, Grundstufe, Komparativ oder Superlativ zu annotieren. Daneben vermisst man jeden Hinweis auf die Person

¹⁶² Online unter: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Lamap-Tagset.pdf>; letzter Zugriff: 31.05.2019.

¹⁶³ Vgl. Helmut Schmid, [Probabilistic Part-of-Speech Tagging Using Decision Trees](#). In: Proceedings of the International Conference on New Methods in Language Processing, (Manchester 1994), 44–49; Web-Präsenz des Annotationstools: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>; letzter Zugriff auf beide URLs: 31.05.2019.

¹⁶⁴ Vgl. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Latin-parameter-file-readme>; letzter Zugriff: 31.05.2019

oder auf die Diathese (aktiv/passiv) des Verbs. Positiv erscheint dagegen die differenzierte Herangehensweise bei der Annotation der verschiedenen Unterarten von Pronomina, wobei, wie bereits gesagt, wiederum die Möglichkeit fehlt, deren Casus zu explizieren.

Die Unterscheidung zwischen Hilfszeitworten und Vollverben kann bei einer gezielten Durchsuchung eines Textcorpus hilfreich sein.

Einzelheiten mag Lamap anderen Tagsets voraus haben, zum Beispiel einen eigenen Tag für nicht lateinische Fremdworte (FW), für Eigennamen (NPR) und die Möglichkeit, Interpunktion zu annotieren (SENT und PUN – dies wurde im Falle des Morpheus Tagsets an die syntaktische Annotation ausgelagert).

Das Beispiel „8.“ zur Veranschaulichung der Funktion des Tags SYM für Symbole („Simboli“) ist schlecht gewählt. SYM wurde dem Tagset der Universal Dependencies¹⁶⁵ – dazu in Kürze – entlehnt. Unter den dortigen Beispielen für diese Kategorie findet sich jedoch weder die Zeichenfolge „8.“ noch eine vergleichbare. Semantisch gesehen weisen die Beispiele für SYM in eine ganz andere Richtung.¹⁶⁶

Zahlworte werden in Lamap durch NUM abgedeckt, wobei hier wiederum die Unterscheidung zwischen Kardinalzahl, Ordnungszahlwort und Distributiva fehlt. Insgesamt wirkt das Tagset Lamap damit höchst unausgegoren. Für die Annotation von Wortarten alleine kann es durchaus empfohlen werden, zumal es dezidiert dafür entwickelt wurde; und es ist in Hinblick auf Wortarten sogar feingliederiger als das Morpheus Tagset, dessen Schwerpunkt mehr auf morphologischer Annotation liegt. In Hinblick auf morphologische Annotation bedürfte Lamap jedoch dringend einiger essentieller Ergänzungen, einiger Änderungen bzw. sollte umstrukturiert werden.

Aus dem Gesagten geht klar hervor: Für eine flächendeckende Annotation von morphologischen Merkmalen ist Lamap derzeit völlig ungeeignet und nicht zu empfehlen, auch wenn einzelne Elemente dafür enthalten sind.

¹⁶⁵ <http://universaldependencies.org/u/pos/index.html>; letzter Zugriff: 31.05.2019.

¹⁶⁶ <http://universaldependencies.org/u/pos/SYM.html>; letzter Zugriff: 31.05.2019.

2.1.2.3) Das PoS-Tagset der Sketch Engine

Für die Kennzeichnung lateinischer Wortarten wurde innerhalb des Tools Sketch Engine ein eigenes, jedoch nicht genuin eigenständiges Tagset entwickelt.¹⁶⁷ Es entstand aus der Zusammenarbeit zwischen Barbara McGillivray und Adam Kilgarriff¹⁶⁸ und zielt speziell auf die Verwendung der genannten Software. Diese verfolgt teilweise kommerzielle Zwecke und stellt ein Produkt der Lexical Computing Ltd. dar.¹⁶⁹ Bei der Sketch Engine handelt es sich um ein web-basiertes Tool zur Erstellung, Organisation und zur Analyse sprachlicher Corpora (u.a. mittels komplexer Suchanfragen).

Es findet vielfach in der Lexikographie Verwendung, vor allem in der englischsprachigen, und ist geeignet, sowohl die grammatische Umgebung wie auch das Kollokationsverhalten einzelner Worte nach Frequenz anzuzeigen,¹⁷⁰ anders gesagt: Welches Wort zeigt innerhalb eines vordefinierten Corpus in welcher Häufigkeit welches grammatikalische Verhalten? Welches Wort tritt in welcher Häufigkeit in welcher lexematischen Umgebung auf? Welche Erscheinungen und Wendungen sind insgesamt typisch für eine Sprache, einen bestimmten Text, einen Autor?¹⁷¹

Entwickelt wurde die Sketch Engine zu Beginn der 2000er-Jahre von Adam Kilgarriff, Pavel Rychlý, Pavel Smrž und David Tugwell.¹⁷² Aus der Kooperation zwischen McGillivray und Kilgarriff entstand bis spätestens 2013 die Sammlung

¹⁶⁷ Online unter: <https://www.sketchengine.eu/latin-part-of-speech-tagset/>; letzter Zugriff: 31.05.2019.

¹⁶⁸ Vgl. Adam Kilgarriff, Barbara McGillivray, Tools for Historical Corpus Research, and a Corpus of Latin. In: Paul Bennett, Martin Durell, Silke Scheible, Richard J. Whitt (Hgg.), New Methods in Historical Corpora (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache – Corpus Linguistics and Interdisciplinary Perspectives on Language Bd. 3, Tübingen 2013), 247–256; online unter: https://www.researchgate.net/publication/236857134_Tools_for_historical_corpus_research_and_a_corpus_of_Latin; letzter Zugriff: 31.05.2019.

¹⁶⁹ Kommerzielle Web-Seite: <https://www.sketchengine.eu/>; letzter Zugriff: 31.05.2019.

¹⁷⁰ Vgl. Kilgarriff, McGillivray, Tools for Historical Corpus Research (wie Anm. 168), 249–251.

¹⁷¹ Vgl. hierzu auch: <https://www.sketchengine.eu/what-can-sketch-engine-do/>; letzter Zugriff: 31.05.2019.

¹⁷² Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, David Tugwell, The Sketch Engine. In: Sandra Vessier, Geoffrey Williams (Hgg.), Proceedings of the 11th EURALEX International Congress (Lorient 2004), 105–115; online unter: <https://euralex.org/publications/the-sketch-engine/>; letzter Zugriff: 31.05.2019; vgl. auch: Adam Kilgarriff, David Tugwell, Word Sketch. Extraction and Display of Significant Collocations for Lexikography. In: Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation (Toulouse 2001), 32–38; online unter: <https://www.kilgarriff.co.uk/Publications/2001-KilgTugwell-ACLcollos-Sketches.pdf>; letzter Zugriff: 31.05.2019.

LatinISE, das erste lateinische Textcorpus, das innerhalb der Sketch Engine verfügbar (wenngleich nicht kostenlos zugänglich) ist.¹⁷³ Dieses enthält nach den ursprünglichen Angaben der Herausgeber rund 13 Millionen sprachlicher Einheiten, welche lemmatisiert und ihrer Wortart bzw. ihrer Funktion nach annotiert wurden. Dabei wird leider nicht ganz klar ersichtlich, ob es sich bei diesen sprachlichen Einheiten bloß um Worte oder um Tokens (also einschließlich Satzzeichen) handelt.¹⁷⁴

Die in LatinISE enthaltenen Texte decken jedenfalls einen äußerst breiten Zeitraum ab – vom 2. Jahrhundert vor Christus bis in das 21. Jahrhundert unserer Zeitrechnung. Sie wurden aus Online-Quellen (!) zusammengetragen, welche „texts from standard editions“¹⁷⁵ in digitalisierter Form (HTML) enthalten: *Lacus Curtius*¹⁷⁶, *IntraText*¹⁷⁷ und *Musisque Deoque*¹⁷⁸.

Betrachtet man die Annotationsergebnisse vor der manuellen Bereinigung¹⁷⁹, wurden diese Texte offenbar mit einem SFST-Tool annotiert (hierzu in Kürze). Zur Weiterentwicklung dieses vielversprechenden Projektes insofern, als ursprünglich geplant war, die Tokens auch um morphologische und um syntaktische Informationen anzureichern, ist es auf Grund des frühzeitigen Todes Kilgarriffs im Frühjahr 2015 bedauerlicherweise nicht gekommen.

Somit befindet sich das Tagset der Sketch Engine noch auf dem Stand des Projektes von 2013. Es stellt eine reduzierte Variante des Tagsets Lamap dar.

¹⁷³ Vgl. <https://www.sketchengine.eu/latin-corpus/>; letzter Zugriff: 31.05.2019.

¹⁷⁴ Vgl. Kilgarriff, McGillivray, *Tools for Historical Corpus Research* (wie Anm. 168), 247–256; in diesem Artikel spricht McGillivray einerseits von 13 Millionen Worten (248), andererseits von 13 Millionen Tokens (255). Hinsichtlich der Anzahl der annotierten sprachlichen Einheiten divergieren diese Angaben auch im Vergleich zu jenen, die dem Web-Auftritt der Sketch Engine zu entnehmen sind. Dort ist (mit Stand 31.05.2019) von 11.036.900 Millionen *Worten* die Rede (vgl. <https://www.sketchengine.eu/corpora-and-languages/corpus-list/>). Es wurden bis zu diesem Datum offenbar Bereinigungen bei der Tokenisierung vorgenommen. Gleichwohl liefert das für die Annotation verwendete Tagset (siehe unten) einen Hinweis: Da es einen Tag für Interpunktion enthält, ist es in Summe wahrscheinlich, dass von 13 Millionen Tokens ausgegangen werden muss. McGillivray, *Methods in Latin Computational Linguistics* (wie Anm. 50), 23 spricht indes wiederum von 13 Millionen Worten.

¹⁷⁵ Kilgarriff, McGillivray, *Tools for Historical Corpus Research* (wie Anm. 168), 251. Die mit diesem Vorgehen verbundene Problematik ist aus editionswissenschaftlicher Perspektive evident: Die editorische Qualität des einzelnen Textes kann bei dieser Art Kompilation eines Online-Corpus beträchtlicher Größe (zumal auf Basis von Online-Quellen!) nur äußerst schwer überprüft werden. Die Computer- bzw. die historisch arbeitende Corpuslinguistik nimmt damit in Kauf, dass auch fehlerbehaftete Editionen in Online-Corpora kanonisiert und dadurch zur Ausgangsbasis weiterer Untersuchungen gemacht werden.

¹⁷⁶ <http://penelope.uchicago.edu/Thayer/E/Roman/home.html>; letzter Zugriff: 31.05.2019.

¹⁷⁷ <http://www.intratext.com>; letzter Zugriff: 31.05.2019.

¹⁷⁸ <http://www.mqdq.it>; letzter Zugriff: 31.05.2019.

¹⁷⁹ Vgl. McGillivray, *Methods in Latin Computational Linguistics* (wie Anm. 50), 24f.

Entsprechend der vorläufigen Zielsetzung von LatinISE (nur PoS-Tagging und Lemmatisierung, keine Explizierung morphologischer Information!) wurde Lamap seiner (wie wir gesehen haben recht unvollständigen) morphologischen Komponenten entkleidet und weiter reduziert. Das PoS-Tagset der Sketch Engine besteht somit nur mehr aus dreizehn (!) Teilen:

TAG	Part of speech	Example
N	noun	salus
V	verb	sapere
PUN	punctuation	,
ADJ	adjective	Gregorii [sic!]
C	conjunction	et
ADV	adverb	forsitan
PRE	preposition	ab
PRO	pronoun	te
P	participle	dictum
NUM	numeral	3, duos
E	eclamation	Ma
NONLAT	nonlatin word	
SUPINE	supine (verb)	petitum

Unter ausdrücklicher Berücksichtigung dessen, dass das Tagset der Sketch Engine vorerst nur Tags für Wortarten bieten möchte, erfüllt es hinlänglich seinen Zweck. So wie jener Teil bei Morpheus, der für Wortarten-Annotation bestimmt ist, könnte es aus linguistischer Sicht jedoch noch beträchtlich erweitert werden. Zumal da es vorerst noch keine morphologischen Tags enthält, trägt das Argument der anzustrebenden Kürze nicht. Gegenwärtig erweist das Tagset der Sketch Engine sich demnach als äußerst reduziert und es zeigen sich ähnliche Mängel wie bei Morpheus, wo jedoch auch morphologische Tags enthalten sind.

Die Unterscheidung zwischen beiordnender und unterordnender Konjunktion fehlt. Es scheint allerdings nicht ausgeschlossen, dass man diese Unterscheidung an die geplanten Treebanks auslagern wollte.

Schwerer wiegt dagegen, dass für die verschiedenen Arten von Pronomina wiederum nur ein Tag zur Verfügung steht (PRO). Zwischen Grundzahlworten, Ordnungszahlworten und Distributiva wird ebenfalls nicht unterschieden.

Postpositionen scheinen wie bei Morpheus nicht auf. Die Unterscheidung zwischen Interjektion und Ausruf, die bei Morpheus noch vorhanden ist, fehlt zwar ebenso, kann aber als verzichtbar angesehen werden; ähnlich die Unterscheidung zwischen Supinum I und Supinum II.

Nicht oder nur schwer verzichtbar erscheinen auf den ersten Blick hingegen eigene Tags für die Unterscheidung der „nd-Formen“, also für Gerundium und Gerundiv. Hierfür fehlen eigene Tags. Beide, sowohl Gerundium als auch Gerundiv, muss man daher unter V (Verb) subsumieren. Dies lässt sich durchaus argumentieren, zumal wenn man Gerundium und Gerundiv nicht als eigene Wortarten ansieht, sondern dem morphologischen Spektrum zuordnet, wie auch im Falle des Morpheus-Tagsets. Grundsätzlich werden Gerundium und Gerundiv ja von Verben abgeleitet. Allerdings wird durch die Subsumierung unter die Kategorie V die Unterscheidung zwischen Gerundium und Gerundiv verwischt: Im einen Falle handelt es sich um ein Verbaladjektiv (Gerundiv), im anderen Fall um ein Verbalsubstantiv neutralen Geschlechtes (Gerundium). Egal, ob als morphologisches Phänomen oder als eigene Wortart – eine eigenständige Kennzeichnung wäre in jedem Fall wünschenswert.

Denn in ähnlicher Weise könnte man beispielsweise die Partizipien unter die Verben subsumieren, was allerdings stark zu hinterfragen ist und was bis jetzt – soweit dem Verfasser der vorliegenden Arbeit bekannt – in keinem lateinischen Tagset so gehandhabt wurde (mit einer Ausnahme – siehe unten!).

Eine Zuordnung der nd-Formen in den syntaktischen Bereich der Treebanks erscheint hingegen nicht sinnvoll.

Zu begrüßen ist im Falle des „Sketch-Tagsets“, dass es eigene Tags für nicht lateinische Fremdworte enthält, was man, wie wir gesehen haben, bei Morpheus vergeblich sucht. Und so, wie im Falle des „Mutter-Tagsets“ Lamap, wird auch Interpunktion annotiert, allerdings ungeachtet dessen, ob es sich um Satz beendende Interpunktion handelt oder um satzinterne. Hier wird bei Lamap genauer unterschieden.

Aus Sicht der Editionswissenschaft ist wiederum das Fehlen eines eigenen Tags für Eigennamen zu bemängeln. Gegenüber Lamap, wo für diesen Zweck das Label NPR – *nomen proprium* – zur Verfügung steht, wurde darauf verzichtet. Das PoS-Tagset der Sketch Engine stellt demnach ein kleinstmögliches Minimum für die Annotation von Wortarten dar, bei dessen Verwendung man einige Abstriche hinsichtlich Genauigkeit hinnehmen muss.

Ungeachtet seiner erwähnten Unzulänglichkeiten im Bereich der Morphologie ist dem Mutter-Tagset Lamap bei der Annotation von Wortarten in jedem Fall der Vorzug zu geben. Im Falle des „Sketch-Tagsets“ sollte jedoch berücksichtigt werden, dass es sich auf Grund des frühzeitigen Todes Kilgarrioffs um ein unvollendetes Projekt handelt.

2.1.2.4) Das morphologische Tagset des Index Thomisticus Online

P	ATT	VAL	C
1	Flexional-Type	Nominal (only degrees and cases)	1
		Participial	2
		Verbal	3
		Invariable	4
		Pseudo-lemma	5
2	Nominals-Degree	Positive	1
		Comparative	2
		Superlative	3
		Not stable composition	4
		None	-
3	Flexional-Category	I decl	A
		II decl	B
		III decl	C
		IV decl	D
		V decl	E
		Regularly irregular decl	F
		Uninflected nominal	G
		I conjug	J
		II conjug	K

		III conjug	L
		IV conjug	M
		Regularly irregular conjug	N
		Invariable	O
		Prepositional (always or not) particle	S
		None	-
4	Mood	Active indicative	A
		Pass/Dep indicative	J
		Active subjunctive	B
		Pass/Dep subjunctive	K
		Active imperative	C
		Pass/Dep imperative	L
		Active participle	D
		Pass/Dep Participle	M
		Active gerund	E
		Passive Gerund	N
		Pass/Dep gerundive	O
		Active supine	G
		Pass/Dep supine	P
		Active infinitive	H
		Pass/Dep infinitive	Q
		None	-
5	Tense	Present	1
		Imperfect	2
		Future	3
		Perfect	4
		Plusperfect	5
		Future perfect	6
		None	-
6	Participials-Degree	Positive	1
		Comparative	2
		Superlative	3
		None	-

7	Case/Number	Singular Nominative	A
		Plural Nominative	J
		Singular Genitive	B
		Plural Genitive	K
		Singular Dative	C
		Plural Dative	L
		Singular Accusative	D
		Plural Accusative	M
		Singular Vocative	E
		Plural Vocative	N
		Singular Ablative	F
		Plural Ablative	O
		Adverbial	G
		Casus “plurimus”	H
		None	-
8	Gender/Number/Person	Masculine	1
		Feminine	2
		Neuter	3
		I singular	4
		II singular	5
		III singular	6
		I plural	7
		II plural	8
		III plural	9
		None	-
9	Composition	Enclytic -ce	A
		Enclytic -cum	C
		Enclytic -met	M
		Enclytic -ne	N
		Enclytic -que	Q
		Enclytic -tenus	T
		Enclytic -ve	V
		Ending homographic with enclytic	H

	Composed with other form	Z
	As lemma	W
	None	-
10	Formal-Variation	
	I variation of wordform	A
	II variation of wordform	B
	III variation of wordform	C
	Author mistake, or bad reading?	X
	None	-
11	Graphical-Variation	
	Baseform	1
	Graphical variations of “1”	2, 3,...
	None	-

Das morphologische Tagset¹⁸⁰ des Index Thomisticus Online¹⁸¹, eines Teils des Corpus Thomisticum¹⁸², stellt innerhalb des hier gezogenen Vergleiches eine Besonderheit dar – zum einen auf Grund seines Umfangs: Es stehen nicht weniger als sechshundneunzig (!) Tags zur Verfügung. Zum anderen fällt auf, dass es fast ausschließlich morphologische Informationen kodiert.

Dies mutet insofern seltsam an, als das Tagset ausdrücklich zur Verwendung in Verbindung mit dem Tree-Tagger zur Verfügung gestellt wird,¹⁸³ der zwar darauf trainiert ist, der aber, wie wir gesehen haben, hauptsächlich dazu dient, Wortarten und Lemmata zu annotieren. Allerdings: Tags zur Annotation von Wortarten sind im Tagset des Index Thomisticus Online streng genommen gar nicht enthalten. Vielmehr werden Wortarten indirekt ausgedrückt – und zwar über die Kategorien Flexional-Type bzw. Flexional-Category – dies allerdings nur sehr ungenau bzw. unvollständig. Und: Nicht immer wird auf Basis des Tagsets klar, wofür ein Tag verwendet wird bzw. wofür er steht. Was ist etwa unter einem „Pseudo-Lemma“ zu verstehen? Beispiele wären hier hilfreich gewesen.

¹⁸⁰ https://itreebank.marginalia.it/doc/Tagset_IT.pdf; letzter Zugriff: 31.05.2019.

¹⁸¹ <http://www.corpusthomisticum.org/it/index.age>; letzter Zugriff: 13.06.2019.

¹⁸² <http://www.corpusthomisticum.org/>; letzter Zugriff: 31.05.2019.

¹⁸³ Vgl. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>; letzter Zugriff: 31.05.2019.

Ob das Tagset seinen Ursprüngen nach auf Padre Roberto Busa S.J. (1913–2011), den Begründer des Index Thomisticus und Wegbereiter der Digital Humanities selbst zurückgeht, konnte im Rahmen der vorliegenden Arbeit nicht nachgewiesen werden.¹⁸⁴

„[T]he Index Thomisticus is considered to be a pathfinder resource in humanities computing and computational linguistics.“¹⁸⁵

Das Tagset stellt seinem Umfang nach den Anspruch, die morphologischen Phänomene der lateinischen Sprache (bei Thomas von Aquin) möglichst feingliedrig und möglichst vollständig abzubilden.

Es ist nicht zu verwechseln mit jenem Tagset, das für die *syntaktische* Annotation der Index Thomisticus *Treebank* (ITTB¹⁸⁶ – nicht identisch mit dem Corpus Thomisticum bzw. mit dem Index Thomisticus Online!) verwendet wurde.¹⁸⁷ Dieser Baumbank liegen wie im Falle der Perseus Digital Library die Guidelines der Prague Dependency Treebank¹⁸⁸ zu Grunde¹⁸⁹. Die dafür verwendeten Primärdaten, also die Texte des Hl. Thomas von Aquin, durchliefen dabei mehrere Annotationsverfahren¹⁹⁰. Im Jahr 2015 wurden sie auch mit den Tags der Universal Dependencies versehen.¹⁹¹

¹⁸⁴ Der Verfasser der vorliegenden Arbeit konnte vorläufig keine gedruckte Publikation Busas ausfindig machen, in der das vorgestellte Tagset enthalten wäre.

¹⁸⁵ <https://itreebank.marginalia.it/view/resources.php>; letzter Zugriff: 31.05.2019.

¹⁸⁶ Vgl. Marco Passarotti, Verso il Lessico Tomistico Biculturale. La Treebank dell Index Thomisticus. In: Diego Femia, Raffaella Petrilli (Hgg.), Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, 14–16 Settembre 2006 (Roma 2007), 187–205; Web-Präsenz des Projektes: <https://itreebank.marginalia.it/view/projet.php>; letzter Zugriff: 31.05.2018;

vgl. ebenso: Marco Passarotti, Leaving Behind the Less-Resourced Status. The Case of Latin through the Experience of the Index Thomisticus Treebank. In: Proceedings of the 7th SaLTMiL Workshop on the Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010 (La Valletta 2010), 27–32; online unter: <http://docplayer.net/7749747-Leaving-behind-the-less-resourced-status-the-case-of-latin-through-the-experience-of-the-index-thomisticus-treebank.html>; letzter Zugriff: 31.05.2019.

¹⁸⁷ Dass auf der Homepage des sprachwissenschaftlichen Seminars der Universität Tübingen (<http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/index-thomisticus-baumbank.html>) auf das hier wiedergegebene morphologische Tagset verwiesen wird (als Tagset der Baumbank! – vgl. http://www.sfs.uni-tuebingen.de/fileadmin/user_upload/ITT-tagset.pdf) scheint nicht korrekt zu sein; letzter Zugriff: 31.05.2019.

¹⁸⁸ <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html>; eine aktuellere Version: <https://ufal.mff.cuni.cz/pdt3.5>; letzter Zugriff auf beide URLs: 31.05.2019.

¹⁸⁹ Vgl. <https://itreebank.marginalia.it/view/ittb.php>; letzter Zugriff: 31.05.2019.

¹⁹⁰ Vgl. https://github.com/UniversalDependencies/UD_Latin-ITTB/tree/master; letzter Zugriff: 31.05.2019.

¹⁹¹ Vgl. <http://universaldependencies.org/treebanks/la-comparison.html>; letzter Zugriff: 31.05.2019.

Über die Benutzeroberfläche des online verfügbaren Index Thomisticus ist das morphologische Tagset in der hier wiedergegebenen Form nicht sichtbar.

Es kodiert linguistische Phänomene wie im Falle des Morpheus-Tagsets in Form von Ziffern und von Buchstaben, die je nach Position ihrer Nennung – es gibt insgesamt elf (!) Positionen – unterschiedliche Bedeutungen annehmen können. Wie im Falle der Perseus Digital Library und wie im Falle der Latin Dependency Treebank werden die Tags auf der Benutzeroberfläche (vgl. Anm. 181) in gängigere Abkürzungen umgewandelt.

Die vordergründig feine morphologische Untergliederung in insgesamt 96 Tags scheint auf den ersten Blick der facettenreichen Sprache des Heiligen Thomas geschuldet.¹⁹² Bei näherer Betrachtung zeigen sich jedoch Defizite bei der Organisation des Tagsets. Anstatt mit Wortarten operiert es als einziges der hier vorgestellten mit den Kategorien Flexionstyp (Flexional-Type) und Flexionskategorie (Flexional-Category). Damit werden Tokens einerseits einem übergeordneten Flexionsschema zugeordnet (= Flexional-Type > Flektiert ein Token wie ein Nomen, ein Partizip oder ein Verb?). Andererseits werden Deklinations- bzw. Konjugationsklassen als solche ausgewiesen (Flexional-Category). Dies stellt unter den hier präsentierten Tagsets ein Alleinstellungsmerkmal dar, was prinzipiell zu begrüßen ist. Was dabei jedoch nicht wenig verwundert: Die Kategorie Flexional-Type enthält neben Nominal, Verbal, Invariable und Pseudo-Lemma zwar einen Tag „Participial“, unterscheidet aber nicht zwischen Präsens-, Perfekt- und Futurpartizipien. Die beiden letztgenannten deklinieren jedoch anders als das Präsenspartizip. Außerdem enthält die Kategorie Flexional-Type keinen eigenen Tag, mit dem Pronomina, Zahlworte oder Adjektive als solche ausgewiesen werden könnten. Das Fehlen dieser Kategorien verwundert insofern, als die Auslagerung von Wortarten in den syntaktischen Bereich der Treebanks keinen Sinn macht, die fehlenden PoS-Tags also auch dort nicht wettgemacht werden können.

Die fehlende Unterscheidung der Partizipien wird immerhin durch die Möglichkeit ausgeglichen, entsprechende Tokens mit Tags der Flexional-Category ihrer Deklinationsklasse zuzuordnen. Adjektive und Zahlworte jedoch müssen

¹⁹² Auf sie kann in der vorliegenden Arbeit in keinsten Weise eingegangen werden. Auch sollen aus dem hier vorgestellten Tagset nur die wichtigsten Charakteristika besprochen werden, sofern sie ein Unterscheidungsmerkmal im Vergleich zu anderen Tagsets darstellen.

gemäß der Kategorie Flexional-Type unter die Nomina (!) subsumiert werden, da sie, vergrößert gesagt, weitestgehend wie diese deklinieren, was in Einzelheiten allerdings nicht befriedigt. Nach diesem Prinzip nämlich hätte man ebenso gut die Partizipien weglassen und unter die Flexion der Nomina subsumieren können, da auch sie, wiederum vergrößert gesprochen, weitestgehend wie diese deklinieren. Und schlicht katastrophal steht es um die Möglichkeit, Pronomina als solche zu annotieren: Besitz anzeigende Fürworte (wie *meus, tuus, suus, noster* und *vester*) muss man ebenfalls den Nomina hinzurechnen (in der Kategorie Flexional-Type), weil sie weitestgehend wie diese abgewandelt werden. Ansonsten lassen sich Possessivpronomina nicht als das ausweisen, was sie sind.

Personal-, Reflexiv-, Interrogativ-, Relativ- und Demonstrativpronomina schließlich kann man weder ihrer Wortart nach als solche kennzeichnen, noch hält das Tagset adäquate Kategorien für ihre jeweils unterschiedlichen Deklinationsmuster bereit. Tags für die Pronominaldeklinatation etwa, nach der auch Worte wie *alter, unus* oder *uterque* abgewandelt werden (im Gegensatz beispielsweise zu den Personal- oder den Reflexivpronomina wie *ego, mei, mihi, ...; nos, nostri, nobis, ...; sui, sibi, se* et c.) fehlen vollständig. Etliche Pronomina kann man demnach nur mit dem Tag „Regularly irregular decl.“ annotieren (aus der Gruppe Flexional-Category) – eine Lösung, die nicht befriedigt.

Es wären tatsächlich *alle* Deklinationen zu berücksichtigen gewesen, wenn es schon die Möglichkeit gibt, diese zu kennzeichnen. Die Gruppe Flexional-Category bietet nicht genügend Tags, um alle Substantivklassen hinreichend abzubilden. Nur fünf werden insgesamt angeboten. Hauptworte der Konsonanten-, der Misch- und der i-Stämme sind dabei unter „III.decl.“ subsumiert. Damit entfällt die Möglichkeit, innerhalb der genannten Muster exakt voneinander zu unterscheiden. Dabei ist es keinesfalls so, dass etwa der Token *turri* (neben *turre* – nicht zu verwechseln mit *ture*) bei Thomas von Aquin nicht belegt wäre, sei dies auf die Überlieferung oder sei dies auf den Aquinaten selbst zurückzuführen.¹⁹³ Sucht man im Index Thomisticus Online (vgl. Anm. 181) übrigens nach dem

¹⁹³ Dass parallel auftretende Lautvarianten im selben Casus wie z. B. *turri* gegenüber *turre, turrim* gegenüber *turrem* im Rahmen des Index Thomisticus Online bloß als „graphic variants“ (!) bezeichnet werden, verweist darauf, dass sprachgeschichtlich komplexe Phänomene wie ähnlich klingende Auslaute und in Folge die Verschmelzung ganzer Deklinationsmuster in adäquater Form thematisiert werden müssten, wenn es bei DH-Projekten um die elektronische Bearbeitung spät- bzw. mittellateinischer Texte geht.

Token *turri*, so wird diese Form ausschließlich (!) als Dativ Singular ausgewiesen (Button „terms“). In der Konkordanz aber werden Fälle angezeigt, bei denen es sich eindeutig um einen Ablativ Singular handelt (Button „concordances“ > mit Anzeige des umgebenden Kontexts). Dasselbe Phänomen tritt auch auf, wenn man etwa nach Tokens wie *crudeli*, *discipulo* oder *dicipuli* sucht. Immerhin – man ist sich des Problems bewusst. Denn in solchen Fällen wird vermerkt: „Homographs within the same lemma entry have not yet been divided“. Die Arbeit am Index Thomisticus Online ist mit Stand Juni 2019 also längst nicht abgeschlossen.

Die Unterscheidung, um welchen Casus es sich im Falle von Homographien handelt, bleibt vorerst dem forschenden Linguisten überlassen, der für diese Unterscheidung entsprechend Zeit investieren muss, um jeden Einzelfall genau zu prüfen. Bei weiterer Betrachtung des Tagsets wirkt der Tag „*prepositional particle*“ für Vorworte etwas eigen. An den Kategorien 4–8 gibt es dem gegenüber wenig auszusetzen. In ihnen sind alle wichtigen morphologischen Merkmale enthalten, nämlich Mood (Modus), Tense, Case/Number sowie Gender, Number, Person. Die Kategorien Case, Number, Person und Gender hätte man dabei auch voneinander trennen und als eigenständige Kategorien etablieren können, so wie bei anderen Tagsets. Ihre Zusammenlegung mutet eigenwillig an.

Der Numerus stellt keine eigenständige Kategorie dar, sondern tritt nur in Verbindung mit Casus oder Person auf. Das führt bei „Number“ zu einer Doppelung und sorgt dafür, dass das Tagset umfangreicher wird als nötig, da jeder Casus und jede Person zweimal, jeweils einmal für den Singular und jeweils einmal für den Plural, mit einem eigenen Tag dargestellt werden muss.

Wo andere Tagsets erwartungsgemäß über zumindest sechs Tags für die üblichen lateinischen Casus (exklusive Lokativ) verfügen und über drei Tags für erste/zweite/dritte Person, besitzt das Tagset des Index Thomisticus nicht weniger als zwölf Tags für die üblichen Casus und sechs Tags für drei Personen.

Die Kategorie der Diathese (aktiv/passiv/Deponens) ist mit den Modi des Verbs (Indikativ, Konjunktiv, Infinitiv, Imperativ) verschmolzen – was wiederum zu einer unnötigen Abundanz führt: Je nach dem, ob aktiv oder passiv/Deponens, ergibt sich bei den Modi des Verbs die Notwendigkeit, einen eigenen Tag zu verwenden. Tags für den Indikativ, für den Konjunktiv, den Infinitiv und den Imperativ sind somit allesamt in doppelter Ausführung vorhanden.

Immerhin: Es wird unterschieden zwischen Gerundiv und Gerundium, was, wie wir gesehen haben, bisher nicht bei allen behandelten Tagsets der Fall war. Die Möglichkeit, den Imperativ II als solchen zu kennzeichnen, fehlt zwar, tritt aber auch in anderen Tagsets selten bis gar nicht auf, zumal es sich um eine sehr rare Form handelt. Schwerer wiegt dagegen die Tatsache, dass ein eigener Tag für die Annotation von Konjunktionen fehlt.

Ausdrücklich berücksichtigt werden im Tagset des Index Thomisticus die Deponentia. Warum die Tags für Grundstufe, Komparativ und Superlativ jedoch doppelt vorhanden sind (nämlich als Nominals-Degree und als Participals-Degree), leuchtet dem Verfasser der vorliegenden Arbeit nicht ein. Wiederum scheint eine unnötige Abundanz vorzuliegen.

Als äußerst positiv ist hingegen das Vorhandensein der Kategorie Nr. 9 hervorzuheben (Composition), in der wichtige Enclitica enthalten sind, die in anderen Tagsets selten so vollständig berücksichtigt werden.

Mit den Kategorien 10 (Formal-Variation) und 11 (Graphical Variation) scheint das Tagset des Index Thomisticus seiner Intention nach über eine rein linguistische Annotation hinausgehen zu wollen – möglicherweise in Richtung einer editorischen Annotation (vgl. den Tag „Author Mistake or bad reading?“).

In Ermangelung aussagekräftiger Beispiele jedoch, wofür die unter 10 angeführten Tags verwendet werden, soll hierauf nicht näher eingegangen werden. Auf die unter Punkt 11 angeführten Tags wurde kurz hingewiesen (vgl. Anm. 193).

Seinen linguistischen Komponenten nach zu urteilen, erweist das Tagset des Index Thomisticus Online sich als unverhältnismäßig und als unnötig groß. Kürzungen wären angebracht und möglich. Gleichzeitig beinhaltet es aus sprachwissenschaftlicher Sicht nicht wenige gravierende Mängel. Wortarten können, wie wir gesehen haben, nur sehr ungenau und sehr unvollständig annotiert werden.

Und obgleich es einzelne linguistische Phänomene genauer expliziert als andere Tagsets, etwa indem es Flexionskategorien als solche ausweist oder Enklitika, vermag es in Summe nicht zu befriedigen.

Eine Empfehlung, das Tagset des Index Thomisticus Online im Rahmen anderer Projekte weiter zu verwenden (und sollten dies auch solche Projekte sein, die explizit mit mittellateinischen Texten christlicher Provenienz und kirchlich-theologischen Inhalts operieren), kann definitiv nicht ausgesprochen werden.

2.1.2.5) Das „Frankfurter Tagset“ und seine Erweiterungen

Die Bezeichnung für das unter diesem Namen vorgestellte Tagset stellt einen Hilfsausdruck dar. Dieser geht auf den Verfasser der vorliegenden Zeilen zurück. Er verweist auf die Herkunft des Tagsets, nicht jedoch auf das DH-Projekt, in dessen Rahmen dieses verwendet wurde. Seitens seiner Urheber, die zum Zeitpunkt seiner Veröffentlichung an der Universität Frankfurt kooperierten, wurde, soweit ersichtlich, keine explizite Benennung vorgenommen. Seine Entwicklung erfolgte im Zuge des Aufbaus eines digitalen Lexikons, des FLL (Frankfurt Latin Lexicon), mit dem primären Zweck, umfangreiche lateinische Textsammlungen zu taggen.¹⁹⁴ Auf den ersten Blick handelt es sich dabei um ein reines PoS-Tagset, um ein Tagset also, das ausschließlich zur Annotation von Wortarten entworfen wurde. Morphologische Elemente sind in der zitierten Literatur kaum bzw. nur indirekt und nicht vollständig greifbar. Die genannte Publikation von Geelhaar, Gleim, Mehler und vor der Brück (vgl. Anm. 194) widmet sich primär der Präsentation des Tools TTLab Latin Tagger (TLT) und der Vorstellung des Frankfurt Latin Lexicon.¹⁹⁵

In einer inhaltlich verwandten Publikation von Eger, Mehler und vor der Brück, ebenfalls aus dem Jahr 2015, sind zwar morphologische Kategorien ersichtlich (case, degree, gender, mood, number, person, tense & voice), allerdings fehlen Angaben darüber, welche Tags konkret *innerhalb* dieser Kategorien, also für die konkrete morphologische Erscheinung, verwendet wurden.¹⁹⁶ Dem Verfasser der vorliegenden Arbeit sind die morphologischen Tags gleichwohl über den

¹⁹⁴ Vgl. Tim Geelhaar, Rüdiger Gleim, Alexander Mehler, Tim vor der Brück, Towards a Network Model of the Coreness of Texts. An Experiment in Classifying Latin Texts using the TTLab Latin Tagger. In: Chris Biemann, Alexander Mehler (Hgg.), Text Mining. From Ontology Learning to Automated Text Processing Applications (Berlin/New York 2015), 87–112; online unter: https://www.researchgate.net/publication/281461116_Towards_a_Network_Model_of_the_Coreness_of_Texts_An_Experiment_in_Classifying_Latin_Texts_Using_the_TTLab_Latin_Tagger; letzter Zugriff: 31.05.2019; dass auf diese Publikation referenziert wird, obgleich das Tagset daraus nicht explizit hervorgeht, geschieht nach ausdrücklicher Rücksprache des Verfassers dieser Zeilen mit Rüdiger Gleim vom Text Technology Lab der Universität Frankfurt und auf dessen ausdrückliches Ersuchen.

¹⁹⁵ Vgl. Geelhaar, Gleim, Mehler, vor der Brück, Towards a Network (wie Anm. 194), 90 ff.; den Informationen von Rüdiger Gleim verdankt der Verfasser der vorliegenden Zeilen die Kenntnis, dass morphologische Elemente von Anfang an integraler Bestandteil des Frankfurter Tagsets waren. Rüdiger Gleim sei an dieser Stelle herzlich für diese Informationen gedankt.

¹⁹⁶ Vgl. Eger, Mehler, vor der Brück, Lexicon-Assisted Tagging and Lemmatization (wie Anm. 145), 106 & 109.

TokenEditor, das Annotations-Tool des Austrian Centre for Digital Humanities der ÖAW in Wien, zugänglich. Auf Basis dessen werden die morphologischen Tags hier in Kürze vorgestellt.

Zunächst jedoch zu den PoS-Tags aus Frankfurt, also zu jenen für die Annotation von Wortarten: Da streng genommen nicht alle, sondern nur die gängigsten Wortarten abgedeckt werden, darf das PoS-Tagset aus Frankfurt als etwas reduziert bezeichnet werden. In jüngerer Vergangenheit wurde, wie gesagt, die Performance mehrerer Annotationstools damit getestet, darunter Lapos, Mate, der TreeTagger, FLORS und MarMot.¹⁹⁷ Es setzt sich aus folgenden PoS-Tags zusammen:

verb: V

adjective: ADJ

adverb: ADV

normal noun: NN

anthroponym: NP

named entity: NE

pronoun: PRO

ordinal number: ORD

cardinal number: NUM

distributive number: DIST

foreign material: FM

conjunction: CON

preposition: AP

interjection: ITJ

non word: XY

particle: PTC

¹⁹⁷ Vgl. Eger, Gleim, Mehler, Lemmatization and Morphological Tagging (wie Anm. 145) sowie Eger, Mehler, vor der Brück, Lexicon-Assisted Tagging and Lemmatization (wie Anm. 145).

non-terminatory punctuation: \$,

terminatory punctuation: \$.

other punctuation: \$(

Wiederum gilt es, bei seiner Evaluation zwei verschränkte Perspektiven zu berücksichtigen, eine editions- und eine sprachwissenschaftliche.

Aus Sicht beider Disziplinen ist die Unterscheidung zwischen normal noun (NN) und Anthroponym (NP) positiv hervorzuheben. Der offen und flexibel gestaltete Tag NE für named entity lässt sich *auch*, aber eben nicht nur für die Annotation von Toponymen verwenden. Ebenfalls zu begrüßen ist die Unterscheidung zwischen ordinal, cardinal und distributive number (ORD/NUM/DIST).

Eine Unterscheidung hingegen zwischen unter- und beiordnenden Konjunktionen fehlt. Beide werden unter CON für conjunction subsumiert. Das hat zur Folge, dass ein mit dem Frankfurter Tagset annotiertes Corpus nicht danach befragt werden kann, wie groß der Anteil der Haupt- bzw. der Nebensätze in ihm ist.

Dass kein eigener PoS-Tag für Partizipien vorhanden ist, liegt daran, dass dieser in den morphologischen Bereich ausgelagert wurde; hierzu in Kürze. Der Tag PTC steht nicht für Partizipien, sondern ausdrücklich für Partikel, wobei nicht näher definiert wird, für welche. Bei der Annotation von Partizipien und auch von Supina verwendet man zunächst den Tag V für Verb, wobei nicht zwischen Voll- und Auxiliarverb unterschieden wird.

Die Subsumierung der Partizipien unter die Kategorie der Verben erscheint insofern gerechtfertigt, als sie sich morphologisch gesehen ja von diesen ableiten, doch wäre auf Grund ihrer Funktionen und auf Grund ihrer Häufigkeit nicht nur im Lateinischen zu fragen, ob ihre Kennzeichnung als eigene Wortart nicht doch sinnvoll wäre. Im vorliegenden Fall müssen Partizipien nämlich mit mindestens zwei Tags versehen werden: V für Verb und PARTICIPLE aus dem morphologischen Teil des Tagsets, wovon V entfallen könnte, würden Partizipien als eigene Wortart geführt. Einen positiven Aspekt stellt der Tag FM für foreign material dar.

Dass Satzzeichen annotiert werden können, ist ebenso zu begrüßen, zumal unterschieden wird zwischen satzbeendender Interpunktion, satzinterner und solcher, der andere Funktionen zukommen.

Die Möglichkeit hingegen, Pronomina anders als mit dem General-Tag PRO zu annotieren, fehlt im ursprünglichen Tagset.

Für die Annotation morphologischer Merkmale werden – ersichtlich nur aus dem TokenEditor! – folgende Kategorien und Tags angeboten:

Casus:	NOMINATIVE
	GENITIVE
	DATIVE
	ACCUSATIVE
	VOCATIVE
	ABLATIVE
	LOCATIVE
	INDECLINEABLE
Numerus:	SINGULAR
	PLURAL
Gender:	COMMON
	FEMININE
	MASCULINE
	NEUTER
Person:	PERSON_1
	PERSON_2
	PERSON_3
Tense:	FUTURE
	IMPERFECT
	PERFECT
	PLUPERFECT
	PRESENT
	FUTURE_PERFECT
Mood:	GERUND
	GERUNDIVE
	IMPERATIVE
	INDICATIVE
	INFINITIVE

	PARTICIPLE
	SUBJUNCTIVE
	SUPINE
Voice:	ACTIVE
	PASSIVE
Degree:	POSITIVE
	COMPARATIVE
	SUPERLATIVE

Die wesentlichsten Teile der lateinischen Morphologie sind damit abgedeckt. Ausdrücklich zu begrüßen ist die Unterscheidung zwischen Gerundium und Gerundiv.

Im Bereich der Casus sind die Tags INDECLINEABLE und LOCATIVE positiv hervorzuheben. Sie kommen nicht in jedem der bisherigen Tag-Inventare vor. Über das Fehlen eines eigenen Tags für den seltenen Imperativ II kann allenfalls hinweggesehen werden.

Es fehlt ein eigener Tag für Verba deponentia. Die Kategorie Voice kennt nur aktiv und passiv. Zwecks linguistischer Annotation der Eckhart-Briefe an die Gebrüder Pez wurden im Rahmen vorliegender Arbeit folgende Tags im TokenEditor ergänzt:

	(PoS-)Tag	Erklärung mit Beispielen
PoS (Type):	ADJ_NE	von Eigennamen abgeleitete Adjektive (z.B. Benedictinus)
	ADJ_NN	hauptsächlich gebrauchte Adjektive (z.B. bona = die Güter)
	APPO	Postposition (z.B. nachgestelltes causa, gratia, tenus, ...)
	CARD	Kardinalzahlen (der vorhandene Tag NUM wurde für indefinite Numeralia wie z.B. tot oder aliquot verwendet)
	CC	koordinierende (= beiordnende) Konjunktion (z.B. ac, atque, et, ...)
	CS	subordinierende (= unterordnende) Konjunktion

		(z.B. cum, ut, ...)
	CLI	Enklitikon (z.B. -ce, -ne, -ve, ...)
	PDEM	Demonstrativpronomen (hic, idem, ille, ipse, is)
	PINDEF	Indefinitpronomen (aliqui/-s, quidam, quisquam, ullus)
	PPERS	Personalpronomina (ego, tu, nos, vos)
	PPOSS	Possessivpronomina (meus, tuus, noster, vester)
	PQUEST	Interrogativpronomina (qui, quis, quid)
	PREFL	Reflexivpronomina (sui/sibi/se)
	PREL	Relativpronomina (qui, quae, quod)
	VCA	Verben, die einen Akt der (verbalen) Kommunikation (communicative act) bezeichnen (z.B. dicere, indicare, citare, vocare, nominare, laudare, ...)
	V_NN	von einem Verbum (Partizip) abgeleitetes Substantiv (z.B. praepositus = der Vorsteher, Abt)
Voice:	Dep	Verbum Deponens
Mood:	INF-ACI	Infinitiv als Bestandteil eines AcI
	INF-NCI	Infinitiv als Bestandteil eines NcI

Einschließlich der hier ergänzten Tags ist das erweiterte Frankfurter Tagset derzeit (Stand: Juni 2019) über den TokenEditor der ÖAW verfügbar. Zusätzlichen Erweiterungen steht es prinzipiell offen. Zu beachten ist in diesem Zusammenhang, dass Tags, die sekundär zum ursprünglichen Frankfurter Tagset hinzukamen, stets manuell zugeordnet werden müssen. Nur auf den *ursprünglichen* Bestand des Frankfurter Tagsets ist das mit dem TokenEditor verknüpfte Tagging-Tool MarMot trainiert. Eine automatische Zuweisung an Tokens ist somit nur im Falle der ursprünglichen Tags möglich. Sekundär hinzugefügte müssen, wie gesagt, stets manuell zugeordnet werden, solange ein Tool nicht darauf trainiert wird. Die praktische Annotation der Eckhart-Briefe an die Gebrüder Pez und die darauf folgenden Corpus-Abfragen haben darüber hinaus gezeigt, dass für speziellere linguistische Forschungsfragen die nachträgliche Hinzunahme weiterer (PoS-)Tags wünschenswert sein könnte, so etwa

PoS (Type): ADV_NN (für hauptwörtlich gebrauchte Adverbien, z.B. *parum pecuniae*),

Casus: ABL_ABS (zusätzlich zu ABLATIVE zur Kennzeichnung von Ablativen, die in Ablativis absolutis auftreten),

ACC_ACI (für Akkusative, die einen ACI einleiten).

Innerhalb der Kategorie Mood ist zu erwägen, den Tag PARTICIPLE für Partizipien zu streichen und stattdessen wie folgt zu differenzieren:

Mood: PART_ACI

[für Partizipien, wenn sie innerhalb eines ACI Teil eines passiven Tempus sind, der Infinitiv *esse* aber elliptisch ausgelassen wurde und der ACI somit nicht über den Infinitiv, sondern ausschließlich über das Partizip kenntlich gemacht werden muss, z.B. in folgendem Satz: *Spes est me, dum adhuc vixero, apud Mellicenses diversurum*. („GE 5“: Johann Georg Eckhart an Bernhard Pez, Hannover 1718-VII-07 bzw. in der Edition: Brief 959).

PART_PC

(für Partizipien, wenn sie Teil eines Participium coniunctum sind)

PART_TEMP bzw. PART_SUBJ

(für Partizipien, wenn sie Teil eines passiven Tempus oder Konjunktivs der Vergangenheit sind).

2.1.2.6) Das Tagset von LatMor

Vom Entwickler des Taggers MarMot und des TreeTaggers, Helmut Schmid, wurde im Jahr 2016 in Kooperation mit Uwe Springmann und Dietmar Najock ein SFST-Tool¹⁹⁸ (kein Tagger im oben skizzierten Sinne!) vorgestellt, „LatMor“¹⁹⁹, welches zu einzelnen, isolierten Zeichenfolgen Informationen generiert, die deren Lemmata und deren morphologische Charakteristika betreffen. Dabei werden auch die Quantitäten der lateinischen Vokale mit berücksichtigt:

„SFST is a toolbox for the implementation of morphological analysers and other tools which are based on finite state transducer technology.“²⁰⁰

„One important application of FSTs is morphological analysis, where a word form such as *translations* might be mapped to the analysis string *translate<V>ion<N>[s]<pl>*.“²⁰¹ Was dies bedeutet, wird in Kürze anhand eines Beispiels expliziert. Zunächst zum Tagset, mit dem LatMor arbeitet. Es besteht aus folgenden Kategorien und Elementen²⁰²:

Person

first	<1>
second	<2>
third	<3>

Gender

feminine	<fem>
masculine	<masc>
neuter	<neut>

¹⁹⁸ Vgl. <http://www.cis.uni-muenchen.de/~schmid/tools/SFST/> bzw. <http://www.cis.uni-muenchen.de/~schmid/tools/SFST/data/SFST-Manual.pdf>; letzter Zugriff auf beide URLs: 31.05.2019.

¹⁹⁹ Vgl. Dietmar Najock, Helmut Schmid, Uwe Springmann, LatMor. A Latin Finite-State Morphology Encoding Vowel Quantity. In: Sabine Erhart (Hg.), Open Linguistics 2/1 (2016), 386–392; online unter: <https://www.degruyter.com/downloadpdf/j/opli.2016.2.issue-1/opli-2016-0019/opli-2016-0019.pdf>;

Online-Präsenz des Tools über die Homepage der Universität München: <http://www.cis.uni-muenchen.de/~schmid/tools/LatMor/>; letzter Zugriff auf beide URLs: 31.05.2019.

²⁰⁰ <http://www.cis.uni-muenchen.de/~schmid/tools/SFST/> (wie Anm. 198); letzter Zugriff: 31.05.2019.

²⁰¹ Helmut Schmid, SFST Manual, o.O., o.D., online unter: <https://www.cis.uni-muenchen.de/~schmid/tools/SFST/data/SFST-Manual.pdf> (wie Anm. 198); letzter Zugriff: 31.05.2019.

²⁰² Vgl. <http://www.cis.uni-muenchen.de/~schmid/tools/LatMor/LatMor-tag-list.txt>; letzter Zugriff: 31.05.2019.

Case

nominative	<nom>
genitive	<gen>
dative	<dat>
accusative	<acc>
ablative	<abl>
vocative	<voc>

Number

singular	<sg>
plural	<pl>

Voice

active	<active>
passive	<passive>
deponens	<deponens>

Part of Speech

adjective	<ADJ>
adverb	<ADV>
conjunction	<CONJ>
interjection	<INTJ>
noun	<N>
numeral	<NUM>
proper name	<PN>
preposition	<PREP>
pronoun	<PRO>
verb	<V>
particle	<PART> (-que, -ne, -ve)

Tense

present	<pres>
future I	<futureI>
future II	<futureII>
future	<future> (imperatives and participles)
imperfect	<imperf>
perfect	<perf>
pluperfect	<pqperf>

Degree

positive	<positive>
comparative	<comparative>
superlative	<superlative>

Mood

indicative	<ind>
subjunctive	<subj>

Number Type

cardinal	<card>
ordinal	<ord>
distributive	<dist>
written with digits	<dig>

Pronoun Types

demonstrative	<dem>
indefinite	<indef>
personal	<pers>
possessive	<poss>
question	<quest>
reflexive	<refl>
relative	<rel>
adjectival	<adj>

Verb Types

supine I	<supinI>
supine II	<supinII>
gerund	<gerund>
gerundive	<gerundivum>
infinitive	<inf>
imperative	<imp>
participle	<part>

Wordform Features

capitalized	<Cap> (such as "Carpe")
all uppercase	<UC> (such as "CAESAR")
wordform with "j"	<i2j> (such as "jus")
alternate form	<alt> (such as "laudavere")

Dieses Tagset stellt einen hinreichend gelungenen Kompromiss zwischen Übersichtlichkeit, Differenziertheit und Genauigkeit dar. Die allermeisten und die wichtigsten Wortarten sowie die wichtigsten morphologischen Phänomene der lateinischen Sprache lassen sich bei seiner Verwendung annotieren. Bei den Casus wäre allenfalls ein Tag für den Lokativ zu ergänzen.

Ausdrückliche Berücksichtigung finden die Verba deponentia unter „Voice“. Die Unterscheidung zwischen Hauptworten und Eigennamen – noun <N> und proper name <PN> – ist aus editionswissenschaftlicher Sicht eindeutig zu begrüßen. Allenfalls könnte man einen eigenen Tag für Toponyme ergänzen. Eine Unterscheidung zwischen unterordnenden und beiordnenden Konjunktionen fehlt indes. LatMor kennt für Bindeworte nur einen Tag: <CONJ>.

Etwas unvorteilhaft wirkt gegenüber dem Tag <part> für Partizipien die Verwendung des Tags <PART> für die Partikel -que, -ne, -ve und -ce, deren Annotierung grundsätzlich zu begrüßen ist. Zur besseren Unterscheidung empfiehlt sich etwa <CLI> oder dergleichen statt <PART>. Innerhalb der Kategorie „Tense“ für die Tempora wird zunächst unterschieden zwischen Futur I <futureI> und Futur II <futureII> für Verben im Indikativ. Hinzu tritt noch der Tag <future> für die spezielle Annotation von Futurpartizipien und von (äußerst seltenen) Futur-Imperativen (Imperativ II; z.B. spectato/spectatote/spectanto), sodass der Tag

<futureI> im Grunde genommen entfallen könnte. Fremdworte, die nicht lateinischen Ursprungs sind, werden in LatMor nicht als solche, sondern allenfalls nach ihrer Wortart annotiert. Auch für Interpunktionen sind keine eigenen Tags vorgesehen. Beides stellt einen gewissen Nachteil dar. Innerhalb der Kategorie „Number Type“ wird zwischen Kardinalzahlen <card>, Ordnungszahlworten <ord> und Distributiva <dist> unterschieden, was zu begrüßen ist. Mit dem Tag <dig> werden Zahlzeichen als solche annotiert. Die Kategorie „Pronoun Types“ erweist sich als erfreulich differenziert und bedarf keiner Ergänzung. In der Kategorie „Verb Types“ sind Tags für beide (!) Supina enthalten. Ebenso wird zwischen Gerundium und Gerundiv unterschieden. Die Kategorie „Wordform Features“ bildet bei LatMor ein Alleinstellungsmerkmal gegenüber anderen Tagsets. Ungeachtet all der genannten Vorteile, die das LatMor-Tagset an sich bietet, gibt es gegenwärtig (Stand: Juni 2019) leider keinen Tagger, der darauf trainiert wäre und der sich mit anderen Tagging-Tools vergleichen ließe. Das SFST-Tool LatMor ist nicht darauf ausgelegt, Texte als solche zu analysieren und zu annotieren, sondern es arbeitet – wie alle SFST-Modelle – strikt auf Basis von Einzelworten. Deren Kontext ist für das Tool irrelevant bzw. nicht fassbar. Jedes Wort bzw. jeder Token wird demnach strikt isoliert betrachtet, und es werden *alle möglichen* Analysen und Varianten dazu generiert, was besonders bei Homographien ein Problem darstellt. Vor allem bei geschlossenen Texten größeren Umfangs kommt es zu einem nicht zu überblickenden Überschuss an Information.

So etwa werden für die Zeichenfolge „malo“ die folgenden Daten (jeweils mit möglichem Grundwort vor den eckigen Klammern) bereitgestellt, deren Fülle sich aus den hier nicht näher verzeichneten Vokalquantitäten ergibt, die im (klassischen) Latein (ursprünglich und streng genommen) bedeutungsunterscheidend waren (mālum = der Apfel; mālum = das Übel; mälle = lieber wollen), ab der Spätantike allerdings nicht mehr:

```
malle<V><pres><ind><active><sg><1>
malum<N><neut><sg><dat>
malum<N><neut><sg><abl>
malus<ADJ><positive><masc><sg><dat>
```

malus<ADJ><positive><masc><sg><abl>
 malus<ADJ><positive><neut><sg><dat>
 malus<ADJ><positive><neut><sg><abl>
 malus<N><masc><sg><dat>
 malus<N><masc><sg><abl>
 malus<N><fem><sg><dat>
 malus<N><fem><sg><abl>

Um das LatMor-Tagset demnach für die automatische linguistische Annotation eines Textes verwenden zu können, müsste zuvor ein manuell annotiertes Trainingscorpus bereitgestellt werden, anhand dessen ein Tagger trainiert würde.

2.1.2.7) Das PoS-Tagset der Universal Dependencies

Das Projekt der Universal Dependencies²⁰³ gehört zu den großen „Flaggschiffen“ der Computerlinguistik der letzten Jahre. Es vereint Ergebnisse zahlreicher Einzelprojekte in sich und speist sich somit aus mehreren Quellen, kann also auf eine lange Entwicklungsgeschichte sowie auf eine außerordentlich breite Forschungsbasis zurückblicken. Es stellt den Anspruch, digitale Infrastrukturen und universelle Tags für die PoS-, die morphologische *und* für die syntaktische Annotation von Texten nahezu jeder beliebigen Sprachen bereitzustellen. Sein Fokus liegt auf dem Treebanking bzw. auf dem Parsing. Gleichwohl klammert es tokenbasierte Annotation von Wortarten nicht gänzlich aus. Gegenwärtig (Stand: Juni 2019) liegen getaggte und geparste Texte für mehr als 70 Sprachen vor (lebende wie historische).²⁰⁴ Seine Wurzeln reichen zurück bis in das Jahr 2005/06.²⁰⁵ Es basiert ursprünglich auf den Stanford Dependencies für Englisch.

²⁰³ Ursprünglich: Claudia Bedini, Núria Bertomeu Castelló, Dipanjan Das, Kuzman Ganchev, Yoav Goldberg, Keith Hall, Jungmee Lee, Ryan McDonald, Joakim Nivre, Yvonne Quirnbach-Brundage, Slav Petrov, Oscar Täckström, Hao Zhang, [Universal Dependency Annotation for Multilingual Parsing](#). In: *Association for Computational Linguistics* (Hg.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Sofia 2013), 92–97;
Webpräsenz des Projektes: <http://universaldependencies.org/introduction.html>; letzter Zugriff: 31.05.2019.

²⁰⁴ Vgl. <http://universaldependencies.org/#language->; letzter Zugriff: 31.05.2019.

²⁰⁵ Vgl. Bill MacCartney, Christopher D. Manning, Marie-Catherine de Marneffe, [Generating Typed Dependency Parses from Phrase Structure Parses](#). In: Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, Daniel Tapias, *European Language Resources Association ELRA* (Hgg.), [Proceedings of the Fifth International Conference on Language Resources and Evaluation \(LREC'06\)](#) (Genua 2006),

Zudem beinhaltet es Elemente der Google Universal Dependencies (2007/2011/2012)²⁰⁶ und des morphosyntaktischen Tagsets Intersect Interlingua²⁰⁷. Jedes dieser Tagsets, keines davon dezidiert oder ausschließlich für die lateinische Sprache entwickelt, weist seinerseits eine längere und eigenständige Entstehungsgeschichte auf. In deren Rahmen erfolgten etliche Anpassungen an spezifische Annotationsvorhaben, Anpassungen, die je nach Beschaffenheit der zu annotierenden Einzelsprache(n) entsprechend variieren konnten.

Die Universal Dependencies stellen, wie angedeutet, den Anspruch, Elemente aus all diesen verschiedenen Tagsets unter einem Großprojekt zu vereinen – mit dem erklärten Ziel, Tags für möglichst alle Sprachen anzubieten. Einzelheiten der Entstehungsgeschichte der Universal Dependencies sind vorläufig auf der Homepage des Projektes samt Literaturangaben dokumentiert (URL: vgl. Anm. 203).

449–454; ferner: Christopher D. Manning, Marie-Catherine de Marneffe, [The Stanford Typed Dependencies Representation](#). In: *Association for Computational Linguistics* (Hg.), Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation (Manchester 2008), 1–20; und:

Samuel R. Bowman, Miriam Connor, Timothy Dozat, Christopher D. Manning, Marie-Catherine de Marneffe, Natalia Silveira, [More Constructions, more genres: Extending Stanford Dependencies](#). In: Proceedings of the Second International Conference on Dependency Linguistics: DepLing; Prague, August 27–30, 2013 (Prag 2013), 187–196; auch: Timothy Dozat, Filip Ginter, Katri Haverinen, Christopher D. Manning, Marie-Catherine de Marneffe, Joakim Nivre, Natalia Silveira, [Universal Stanford Dependencies: A cross-linguistic typology](#). In: *European Language Resources Association ELRA* (Hg.), [Proceedings of the Ninth International Conference on Language Resources and Evaluation \(LREC-2014\)](#) (Reykjavik 2014), 4585–4592.

²⁰⁶ Vgl. Ryan McDonald, Joakim Nivre, [Characterizing the Errors of Data-Driven Dependency Parsing Models](#). In: *Association for Computational Linguistics* (Hg.), Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (Prag 2007), 122–131; ebenso: Dipanjan Das, Slav Petrov, [Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections](#). In: *Association for Computational Linguistics* (Hg.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Portland/Oregon 2011), 600–609; Dipanjan Das, Ryan McDonald, Slav Petrov, [A Universal Part-of-Speech Tagset](#). In: *European Language Resources Association ELRA* (Hg.), [Proceedings of the Eighth International Conference on Language Resources and Evaluation \(LREC-2012\)](#) (Istanbul 2012), 2089–2096.

²⁰⁷ Daniel Zeman, [Reusable Tagset Conversion Using Tagset Drivers](#). In: *European Language Resources Association ELRA* (Hg.), Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May–1 June 2008 (Marrakech 2008), 213–218; ferner: Philip Resnik, Daniel Zeman, [Cross-Language Parser Adaptation between Related Languages](#). In: *Asian Federation of Natural Language Processing* (Hg.), Proceedings of IJCNLP 2008 Workshop on NLP for Less Privileged Languages (Hyderabad 2008), 35–42.

Die Fülle der damit verbundenen Publikationen würde durchaus einen eigenen, einheitlichen und zusammenfassenden Überblick zur Forschungs- und zur Projektgeschichte in gedruckter Form rechtfertigen. Ein solcher ist dem Verfasser der vorliegenden Arbeit gegenwärtig nicht bekannt, kann an dieser Stelle jedoch nicht geboten werden. Die im Rahmen der Universal Dependencies angebotenen PoS-Tags sind größtenteils selbsterklärend:

ADJ: adjective
ADP: adposition
ADV: adverb
AUX: auxiliary
CCONJ: coordinating conjunction
DET: determiner
INTJ: interjection
NOUN: noun
NUM: numeral
PART: particle
PRON: pronoun
PROPN: proper noun
PUNCT: punctuation
SCONJ: subordinating conjunction
SYM: symbol
VERB: verb
X: other

Der Fokus der Universal Dependencies liegt wie erwähnt im Bereich des Treebankings und des Parsings. PoS-Tags spielen eine verhältnismäßig untergeordnete Rolle im Vergleich zu den Tags für die syntaktische Annotation. Entsprechend erweist das PoS-Tagset der Universal Dependencies sich als relativ reduziert, denn wo möglich, wurden Tags in den morphologisch-syntaktischen Bereich der Treebanks ausgelagert. So ist es zu erklären, dass sich unter den Tags für Wortarten kein eigener für Partizipien findet.

Der Tag AUX (auxiliary) steht für Hilfszeitworte im Gegensatz zu Vollverben.

Die Kategorie DET versammelt Artikel und Pronominaladjektive: „Determiners are words that modify nouns or noun phrases and express the reference of the noun phrase in context.“²⁰⁸ Teils handelt es hierbei um Worte, die traditionellerweise unter die Pronomina subsumiert würden (Demonstrativ-, Possessiv-, Indefinit-, Interrogativ- und Relativpronomina: this, my, all, which).

Im Rahmen dreier bekannter Projekte (Proiel, Perseus und Index Thomisticus Treebank) wurden diese Tags auch auf lateinische Texte angewandt. Dabei zeigt sich, dass die PoS-Tags der Universal Dependencies eher als grobe Richtlinie denn als verbindliche Norm angesehen werden. Die Verwendung etwa von AUX gegenüber VERB sowie von PRON gegenüber DET weicht in jedem der genannten Projekte mehr oder weniger stark voneinander ab.²⁰⁹

Anders, als man erwarten würde, hat die Verwendung der UD-Tags bei Proiel, Perseus und Index Thomisticus gerade nicht dazu geführt, dass alle drei Corpora gut untereinander vergleichbar wären.

2.1.3) Zusammenfassung, Desiderate, Empfehlungen

Die laufende Beschäftigung mit digitalen Ressourcen für maschinelle Sprachverarbeitung lateinischer Texte hat bis dato folgendes gezeigt:

(1) Die computerlinguistische Forschung der letzten Jahre hat für literarische (!) Quellen lateinischer Sprache zahlreiche Komplettlösungen erarbeitet – in Form von Tagsets und in Form trainierter Annotationstools. Diese Tools erzielen bei automatischer Annotation teils beachtliche Trefferquoten (im Falle jüngerer Tools: weit über 80, ja sogar leicht über 90% – vgl. Anm. 145–146).

Trotzdem wird empfohlen, im Rahmen von geschichtswissenschaftlichen Editionsunternehmen den Einsatz fertiger Lösungen (also fixer Kombinationen bestehender Tagsets mit bestehenden Tools) genau abzuwägen, speziell dann, wenn es um Textausgaben pragmatischer Schriftlichkeit des Mittelalters und der frühen Neuzeit geht.

Denn: (2) Selbst bei großen und bei von Altphilologen mitbetreuten Forschungsprojekten der Digital Humanities (wie etwa im Falle der Perseus Digital Library) stellt die Qualität der verwendeten PoS-Tagsets zuweilen einen echten Schwach-

²⁰⁸ <http://universaldependencies.org/u/pos/DET.html>; letzter Zugriff: 31.05.2019.

²⁰⁹ <http://universaldependencies.org/treebanks/la-comparison.html>; letzter Zugriff: 31.05.2019.

punkt dar. Partiell ist dies dadurch zu erklären, dass das Hauptaugenmerk etlicher Projekte häufig im Bereich des Parsings und des Treebankings lag und liegt. Bestehende PoS-Tagsets für lateinische Texte wurden zudem fast ausschließlich zum Gebrauch an bereits edierten literarischen Quellen entwickelt, die nachträglich digitalisiert wurden, unter Heranziehung hauptsächlich literarischer Werke der Antike und des Mittelalters.

Pragmatische Quellen, zu denen hier auch nicht-literarische Gebrauchsbriefe, etwa von Gelehrten des 18. Jahrhunderts, gezählt werden, weichen in ihrer Sprache, vor allem in der Lexik, partiell vom hohen Standard literarischer bzw. wissenschaftlich-gedruckter Schriftlichkeit ab und folgen zum Teil eigenen (sozio-)linguistischen Regeln. Es ist also nicht auszuschließen, dass selbst moderne Tagger bei pragmatischer Schriftlichkeit andere (i.e.: geringfügig schlechtere?) Trefferquoten erzielen als bei literarischen Texten, auf deren Tagging sie grundsätzlich trainiert wurden.

Aus Sicht des Diplomatikers etwa erscheint diese Annahme deshalb nicht unplausibel, weil das Latein mittelalterlicher Urkunden bekanntlich eine „völlig andere Welt“ darstellt im Vergleich mit klassisch-lateinischer Literatur, aber teils auch im Vergleich zu literarischen Quellen des Mittelalters, die sprachlich höchst heterogen gestaltet sein können.

Indes gibt es, soweit dem Verfasser vorliegender Zeilen bekannt, gegenwärtig (Stand: Juni 2019) keine digitalen Referenzcorpora, die ausschließlich Quellen pragmatischer Schriftlichkeit des Mittelalters und/oder der frühen Neuzeit enthielten. Dem Vergleich gängiger Tagger hinsichtlich ihrer Performance bei pragmatischen Texten einerseits und bei literarischen andererseits müsste zunächst die Erstellung eines pragmatischen Corpus vorangehen – oder zumindest die Erstellung einer begrenzten Auswahl an entsprechenden Texten. Denn der Begriff „pragmatische Schriftlichkeit“ ist bekanntlich extrem weit gefasst und beinhaltet große Mengen höchst unterschiedlicher Quellen, sodass hier eine starke Vorauswahl getroffen werden müsste. Die Erstellung eines Referenzcorpus lateinischer Urkunden etwa ist nach wie vor ein Desiderat.

(3) In Summe lässt sich beobachten, dass die oftmalige Schwerpunktsetzung bisheriger Großprojekte im Bereich der überaus arbeitsaufwendigen Treebanks teils zum Nachteil jener Tagsets geschah, die parallel zu den Treebanks für die

PoS- und für die morphologische Annotierung konzipiert wurden. Wer neben Wortarten auch morphosyntaktische Merkmale digital explizieren, sich dabei aber nicht der Mühe unterziehen will, Treebanks zu erstellen (aus Zeit- und/oder aus Kostengründen), der findet seine Möglichkeiten im Grunde genommen auf eine sehr kleine Anzahl wirklich brauchbarer Tagsets für die lateinische Sprache beschränkt, Tagsets, die – aus latinistischer Sicht – in den allermeisten Fällen mancher grundlegender Ergänzungen bedürfen.

Neben beträchtlicher Arbeitszeit verlangen Treebanks ihrerseits fundierte Kenntnisse und einige Routine im Umgang mit Abhängigkeitsgrammatiken. Diese sollen nicht Gegenstand der vorliegenden Darstellung sein. Einen entscheidenden Vorteil bieten Treebanks jedoch sehr wohl, nämlich dass die damit syntaktisch annotierten Texte gezielt nach Größen wie Subjekt, Objekt, Prädikat u. dgl. durchsucht werden können.

Insofern – und im Sinne der Praktikabilität innerhalb der Editionswissenschaft – empfiehlt der Verfasser vorliegender Arbeit als Abschluss dieses Kapitels nachdrücklich, bei der Erstellung und/oder bei der Bearbeitung bestehender Tagsets für die lateinische Sprache auch solche Tags in Erwägung zu ziehen, die syntaktische Informationen und Zusammenhänge wenigstens partiell codieren, z.B. Subjekte, Prädikate, Objekte et c., Informationen also, die innerhalb der Computerlinguistik bisher fast ausschließlich mittels Treebanks codiert wurden. Konkrete Vorschläge in dieser Hinsicht werden nach Abschluss vorliegender Arbeit erstellt.

2.2) Tokens, Kontext und Semantik

Es entspricht einem bereits mehrfach wahrgenommenen Trend innerhalb der Digital Humanities, dass auch die Ergebnisse linguistischer Annotation sich gut für die Darstellung in Form von Tabellen und/oder für die Darstellung in Form einer Reihe von (Beleg-)Zahlen eignen (quantitativer Aspekt). Unter Zuhilfenahme etwa der erwähnten Sketch Engine lassen sich innerhalb eines annotierten Textcorpus Frequenz, Belegstellen und Kollokationsverhalten eines gesuchten Lemmas leicht ermitteln und anzeigen, mit verhältnismäßig geringem Aufwand. Auch der TokenEditor (siehe unten) bietet hierfür einfache Möglichkeiten zur Anzeige von Wortfrequenzen und zur Anzeige des sprachlichen Kontexts gesuchter Worte. Das vollständig annotierte Corpus wird, wie eingangs gesagt, zur

möglichst lückenlosen Konkordanz seiner selbst. Nun ist aber die rein quantitative, listenartige Aufzählung von Belegstellen alleine nur von mäßigem Interesse, wenn man bei der Auswertung nicht weiter geht, indem man (a) differenziertere Forschungsfragen hinterlegt und indem man (b) den linguistischen Kontext des einzelnen Tokens ebenfalls mit berücksichtigt. Denn dass eine rein quantitative, listenartige Darstellungsweise tokenzentrierter Annotationsergebnisse der Semantik von Worten alleine nicht gerecht wird, ist evident, zumal bei Vernachlässigung des sprachlichen Zusammenhangs.

Differenziertere Forschungsfragen über bloße Wortfrequenzen und den einzelnen Token hinaus können linguistischer Natur sein, beispielsweise mit diachron-historischem Aspekt und sie können Vergleiche implizieren: Liegen mehrere Corpora in ein und derselben Sprache vor und decken dabei einen bestimmten Zeitraum ab, so können diachrone, corpusbasierte Studien zur Lexik dieser Textsammlungen, zur Frequenz und/oder zur Semantik einzelner Worte bzw. zu einzelnen Satzkonstruktionen durchgeführt werden – mittels digitaler Suchabfragen. Dies wurde eingangs bereits skizziert. Ebenso kann das Kollokationsverhalten von Worten gezielt untersucht werden. Es soll allerdings nicht verschwiegen werden, dass dergleichen Forschungsfragen im Rahmen vorliegender Studie nur in sehr einfacher Form gestellt und beantwortet werden können – vorerst nämlich ohne den Einsatz spezifischer CQL (Context Query Language) und ohne den Einsatz von X-Path.

Unabhängig davon ist bei Fragen nach historisch-vergleichender Semantik ohne Zweifel ein Faktor von entscheidender Bedeutung: der sprachliche Kontext des einzelnen Tokens. Dieser Kontext prägt und bestimmt die Bedeutung von Worten, wie oben skizziert, bekanntlich wesentlich mit, ja bildet *den* wesentlichen semantischen Faktor überhaupt. Dass die Bedeutung von Worten durch deren Gebrauch in der Sprache determiniert ist und gemeinsam mit diesem diachronen Veränderungen unterliegt, bedarf keines gesonderten Hinweises.

Dies im Auge zu behalten, ist für den Zusammenhang der vorliegenden Studie deshalb von essentieller Bedeutung, weil dadurch verdeutlicht wird, dass historische Semantik sich gerade *nicht* oder nur sehr begrenzt auf der Ebene der einzelnen, isolierten sprachlichen Tokens abspielt.

Die digitale Annotation von Wortarten und von morphologischen Phänomenen zielt aber *scheinbar exakt darauf* ab, gerade wenn man auf Treebanks verzichtet: auf den einzelnen, isolierten Token – die kleinste distinkte Einheit des tokenisierten Textes. Die Eigenschaften des Einzeltokens bestimmten die Arbeitsabläufe der linguistischen Annotation in der Regel sehr stark. Allerdings ergeben sich die entscheidende Eigenschaften des Einzeltokens gerade auch in Zusammenwirken mit seinem sprachlichen Kontext.

Daraus wird klar ersichtlich: Texte, die tokenbasiert annotiert wurden, bedürfen nach wie vor der hermeneutisch-interpretativen Auswertung. Annotation liefert hierfür nur eine Hilfestellung. Das sollte gerade im Kontext digitaler Textbearbeitung nicht übersehen werden.

Sollen die Ergebnisse linguistischer Annotation daher nicht auf bloße Wortfrequenzlisten beschränkt bleiben, müssen Forschungsfragen teilweise getrennt, teilweise in Zusammenwirken mit dem tokenbasierten Arbeitsprozess der Annotation selbst formuliert werden – getrennt deshalb, weil semantische Aspekte, wie gesagt, über den einzelnen Token hinaus gehen und dessen Kontext berücksichtigt werden muss, in Zusammenwirken deshalb weil die Möglichkeiten digitaler Suchabfragen sich zunächst (!) auf den einzelnen Token beziehen (wie hier in ihrer einfachsten Form).

Hierzu ein Beispiel: Die Zeichenkette „ut“ ist in so gut wie jedem lateinischen Textcorpus anzutreffen, bezogen auf ihre Frequenz also ein erwartungsgemäß ergiebiges Objekt für tokenbasierte Suchabfragen. Wie häufig dieser Token allerdings in mindestens zwei unterschiedlichen lateinischen Textcorpora vergleichbarer Art und Größe auftritt, ist für sich genommen nur von untergeordnetem Interesse. Denn die Frage nach der Häufigkeit seines Auftretens alleine berücksichtigt nicht, dass er je nach Kontext unterschiedliche Bedeutungen annehmen kann, je nach dem, ob er in Verbindung mit Indikativ oder mit Konjunktiv gebraucht wird, als Einleitungswort oder (unklassisch) im Rahmen eines Vergleiches.

Die Aussage, im Textcorpus X sei die Zeichenkette „ut“ häufiger vorhanden als im Textcorpus Y, kann also nur dann von Belang sein, wenn das semantische Potenzial, die semantische Ambiguität und der mögliche Kontext des Tokens bei der vorangehenden Forschungsfrage ausdrücklich berücksichtigt werden. Nicht

jedes „ut“ ist seiner Bedeutung nach mit jedem anderen vergleichbar. Es gilt demnach zu fragen, inwiefern im Zuge der tokenbasierten linguistischen Annotation bereits der Semantik gemäß disambiguiert werden sollte. Indes scheinen Einzeltoken und Semantik zwei geradezu gegensätzliche Pole zu bilden, die sehr weit auseinander liegen. Semantische Annotation wird auf Tokenebene daher nur sehr schwer durchzuführen sein. Semantische Disambiguierung, etwa von Homographien, ist daher spätestens bei der Abfrage und bei der Interpretation der Annotationsergebnisse vorzunehmen – im Rahmen eines hermeneutisch-interpretativen Prozesses, den digitale Technologien sehr wohl unterstützen, (vorerst) aber nicht gänzlich ersetzen können. Linguistisch differenzierte Forschungsfragen wird man also nur in den seltensten Fällen auf Grund tokenbasierter sprachlicher Annotation alleine beantworten können.

Obiges Beispiel mit „ut“ soll verdeutlichen, dass linguistische Forschungsfragen teilweise getrennt von der tokenbasierten Annotation formuliert werden müssen. Gleichwohl kann eine rein frequenz- und tokenbasierte Suchabfrage nach „ut“ differenziertere Fragestellungen unterstützen. Der Zugriff auf den sprachlichen Kontext des Einzeltokens ist mit digitalen Werkzeugen wie dem TokenEditor (siehe unten) oder mit der Sketch Engine (siehe oben) möglich bzw. wird wesentlich erleichtert.

Die Chance, dass Belege für „ut“ oder „uti“ im Zuge der automatischen Annotation falsch lemmatisiert wurden, ist relativ gering, da der Unterschied zwischen Token und Type (nahezu) entfällt – nur die seltene Nebenform „uti“ ist in dieser Hinsicht ambig: Sie kann Infinitiv Präsens zu „utor“ (= gebrauchen) sein. Zu erwarten ist also, dass eine digitale Suchabfrage nach „ut“ in einem automatisch lemmatisierten Corpus so gut wie alle Ergebnisse korrekt anzeigt, ohne dass dafür manuelle Korrekturen in größerem Ausmaß nötig wären. Insofern wurde das Beispiel „ut“ hier bewusst gewählt – unter Berücksichtigung der Möglichkeiten *tokenbasierter* Annotation.

Ist – wie im TokenEditor und/oder in der Sketch Engine – die Möglichkeit gegeben, den sprachlichen Kontext parallel zum Einzeltoken anzuzeigen, so kann etwa folgende Forschungsfrage relativ schnell beantwortet werden: Gebraucht ein Autor X Tokens wie „ut“ oder „cum“ bevorzugt (oder ausschließlich) mit Indikativ oder mit Konjunktiv? Diese und ähnliche Fragen verknüpfen Semantik

mit Wortfrequenz. Auf ihrer Grundlage lassen sich nach erfolgter linguistischer Basisannotation Aussagen zu lexematisch-semantischen Vorlieben eines Autors treffen. Auch Analysen zur bevorzugten Verwendung von Hypo- bzw. von Parataxe sind möglich, zum Beispiel: Treten beordnende Bindeworte wie „et“ bei einem Autor häufiger auf als unterordnende?²¹⁰ Der schnelle Zugriff auf den sprachlichen Kontext des Einzeltokens erübrigt semantische Annotation auf Tokenebene also mitunter.

Ein zusätzliches Beispiel aus der Praxis soll dazu dienen, eine weitere Problematik der tokenbasierten linguistischen Annotation besser zu verdeutlichen, nämlich den Umgang mit mehrteiligen Eigennamen:

Eine Zeichenkette wie „Carolus“ oder „Alexander“ gewinnt ihre Bedeutung häufig nicht alleine aus sich selbst, zumal dann, wenn daneben eine Zeichenkette wie „Magnus“ aufscheint, die Zeichenkette „Johannes“ ebenso wenig, wenn sich ihre vollständige Bedeutung erst gemeinsam mit den folgenden Tokens ergibt: „Georgius Eckhardus“. Dass erst zwei oder mehrere Tokens zusammen adäquat auf einen Namensträger/eine Namensträgerin (oder auch: auf die Konvention, sie/ihn so zu nennen) verweisen können – das ist zwar hinlänglich bekannt, wird in der streng tokenbasierten Annotationsweise vorerst aber ignoriert, was weiter oben bereits kurz angesprochen wurde (vgl. Anm. 85). Dabei ist es irrelevant, ob es sich bei der Entität, die einen Eigennamen trägt, um eine Person, einen Ort oder um eine Organisation handelt. Das soeben anhand zweier Personennamen skizzierte Problem ergibt sich auch bei Entitäten wie *Terra Sancta* (das Hl. Land) oder *monasterium Sanctae Crucis* – womit vorläufig nur angedeutet werden soll, dass die digitale Prozessierung von *named entities* innerhalb der DH einen eigenen Problem- bzw. Forschungsbereich darstellt, der über tokenbasierte Annotation deutlich hinausgeht.

Ähnliches gilt nicht nur im Falle der lateinischen Sprache bei morphologischer Annotation synthetisch gebildeter Verbformen. Mutatis mutandis sind diese häufig aus mehreren Tokens zusammengesetzt. Man denke hierbei an passive Tempora der Vergangenheit, an den passiven Infinitiv der Vergangenheit oder auch an

²¹⁰ Das bereits erwähnte Potential der linguistischen Annotation soll mit Fragen wie diesen jedoch nur angedeutet werden. Die Praxis der vergleichenden linguistischen Auswertung mehrerer annotierter Corpora wird aus der vorliegenden Arbeit zum größten Teil bewusst ausgelagert (aus Platz- und arbeitsökonomischen Gründen).

synthetisch gebildete, passive Konjunktivformen, die im Lateinischen allesamt aus mehreren Tokens gebildet werden, einem Partizip Perfekt und einer (un-)flektierten Form von *esse*. Erst zwei Tokens zusammen ergeben gemeinsam ein Tempus bzw. einen Modus.

Dasselbe Problem stellt sich auch bei synthetisch gebildeten Steigerungsformen von Adjektiven. In der Anredeformel *Vir maxime reverende* muss *reverende* tokenbasiert als Grundstufe getaggt werden, obwohl semantisch gesehen durch die Verbindung mit *maxime* ein Superlativ vorliegt.²¹¹ Diesen Phänomenen kann eine rein tokenbasierte Annotation nicht gerecht werden, auch und gerade dann, wenn sie unter Verwendung des TokenEditors auf morphologische Merkmale abzielt.

Die Annotationspraxis im Rahmen eines Projektes am ACDH der ÖAW in Wien, an dem der Verfasser vorliegender Zeilen mitwirken konnte, hat zwar gezeigt, dass sich auch in den TokenEditor Funktionen (in XML: Attribute) implementieren lassen, um die skizzierten Probleme einwandfrei zu bewältigen.²¹² Gleichwohl wurden bei besagtem Projekt „nur“ PoS-Tagging (Wortarten-annotation) und Lemmatisierung durchgeführt. Auf morphologische Annotation wurde verzichtet. Daher stellte sich oben skizziertes Problem mehrteiliger Tokens zwar nicht im Falle von Tempora und Modi, wohl aber im Falle von *named entities* bzw. Eigennamen.

2.3) Das Werkzeug: Der TokenEditor

Im Rahmen des bereits mehrfach erwähnten Abac:us-Projektes (vgl. Anm. 9 und 92) am ACDH der ÖAW (damals Institut für Corpuslinguistik und Texttechnologie) wurde der TokenEditor²¹³ entwickelt, um „automatisch [durch Tagger] generierte [Annotationsergebnisse] systematisch [zu] überprüfen [...]“.²¹⁴

Als webbasiertes Tool ermöglicht er (nach passwortgeschütztem Log-In) ortsunabhängig den Zugriff auf automatisch vorannotierte Texte und gestattet so die manuelle Verbesserung maschineller Annotationsergebnisse – „a web

²¹¹ Warum Johann Georg Eckhart an Stelle der verwendeten Formel *Vir maxime reverende* nicht auf *Vir reverendissime* zurückgreift, wäre eigens zu untersuchen.

²¹² Projekt „Totenkult und Jenseitsvorsorge in Wien: Barocke Bruderschaftsdrucke als Forschungsgegenstand der digitalen Geisteswissenschaften“ – Projekt-Nr.: LW 10240 I-III; Leitung: Dr. Claudia Resch.

²¹³ <https://clarin.oeaw.ac.at/tokenEditor/>; letzter Zugriff: 31.05.2019.

²¹⁴ Czeitschner, Resch, Austrian Baroque Corpus (wie Anm. 9), 48.

application for manual annotation (or manual review of automatic annotations) of text. The application is built with AngularJS, PHP, and PostgreSQL and uses a grid module for data manipulation as well as chart modules for visualization. Changes are stored continuously in the database on token level.“²¹⁵

Wortarten-Annotation (PoS-Tagging) sowie Lemmatisierung sind die Hauptanwendungsgebiete des TokenEditors. Grundsätzlich besteht auch die Möglichkeit zur Erweiterung bzw. zur Veränderung des Annotationsspektrums (z.B. für morphologisches Tagging neben Wortarten-Annotation und Lemmatisierung oder in Richtung speziellerer Annotation mehrteiliger Named Entities).

Welche Grundkonfiguration der/die Benutzer*in des TokenEditors vor Beginn der eigentlichen Annotationsarbeit vorfindet, hängt von den individuellen Annotationsvorhaben bzw. -bedürfnissen ab. Es empfiehlt sich, diese vor Beginn der eigentlichen Annotation mit dem Entwickler*innen-Team am ACDH abzusprechen. Der TokenEditor wird dann entsprechend eingerichtet. Werden während des Arbeitsprozesses zusätzliche Tags oder neue bzw. veränderte Annotationskategorien benötigt, so werden diese ebenfalls seitens des ACDH konfiguriert.

Die graphische Benutzeroberfläche des TokenEditor hat die Form einer mehrspaltigen Tabelle. Anzahl und Inhalt der Spalten können je nach Annotationsziel und -bedarf variieren. Die Zeilen der Tabelle ergeben sich aus den Bestandteilen des tokenisierten und in vertikaler Form dargestellten Textes, der annotiert werden soll, also aus dessen Worten und aus dessen Satzzeichen. Den in dieser Form dargestellten Text liest man für gewöhnlich in waagrechten Zeilen von oben nach unten, pro Token eine Zeile. Jeder Wort-Token und jedes Satzzeichen wird mit einer eigenen und einmaligen Token-ID versehen (fortlaufende Nummer beginnend mit 1).

Im Hintergrund der graphischen Benutzeroberfläche wird mit dem Dateiformat XML operiert, in dessen Form die bearbeiteten Daten gespeichert werden. Die graphische Benutzeroberfläche des TokenEditor ermöglicht die Annotation von Texten, ohne seitens der Annotatorinnen und Annotatoren XML-Kenntnisse vorauszusetzen.

²¹⁵ <https://www.oeaw.ac.at/acdh/tools/tokeneditor/>; letzter Zugriff: 31.05.2019.

Der TokenEditor ist über die gängigsten Internet-Browser erreichbar. Für die Bearbeitung von Texten bedarf es einerseits des passwortgeschützten Log-Ins in das Tool selbst, andererseits der Freischaltung der Texte. Die Benützung des TokenEditors ist kostenlos, muss aber in Absprache mit dem Team des ACDH vorgenommen werden.

Nach erfolgreichem Log-In erscheinen in einem eigenen Feld („Your documents“) jene Dateien, für deren Bearbeitung (Annotation) man als Benutzer*in freigeschaltet wurde. Durch das Öffnen (Anklicken) einer Datei im Feld „Your documents“ wird der tokenisierte Text vertikal in Tabellenform angezeigt (Lese-richtung, wie gesagt, von oben nach unten). In der Regel werden 25 Zeilen (Tokens) pro Seite angezeigt. Man kann auf bis zu 1000 Zeilen (Tokens) pro Seite erweitern, wobei dann der Scroll-Balken am rechten Rand des Fensters zu benutzen ist, da in diesem Fall nicht alle Tokens gleichzeitig sichtbar sind.

Jene beiden Spalten, in denen Token-ID und der einzelne Token selbst angezeigt werden, können von Benutzer*innen inhaltlich nicht verändert, wohl aber ein- und ausgeblendet werden. Ebenso lässt die Spaltenbreite sich manuell einstellen. Die beiden letztgenannten Punkte gelten auch für alle anderen Spalten, in denen zunächst die Ergebnisse der automatisch generierten Annotation sichtbar sind (zumeist Lemma und Wortart, optional: morphologische Informationen, z.B. Genus, Casus, Numerus et c., jeweils in einer eigenen Spalte wiedergegeben, auch in Abhängigkeit vom verwendeten Tagset). Mit Ausnahme der ersten beiden Spalten (Token-ID und Token) ist der Inhalt aller weiteren Spalten, wie gesagt, variabel und durch Benutzer*innen (zumeist anhand selbstgesetzter Annotationsrichtlinien) veränderbar, je nach dem, mit welchen (zusätzlichen) Informationen die Tokens versehen werden sollen, und je nach dem, was und wie annotiert werden soll (z.B. Verweise auf Referenzwörterbücher, aus denen die Lemmata entnommen wurden, oder Beginn und Ende zusammengehöriger, mehrteiliger Tokens wie etwa im Falle mehrteiliger Eigennamen et c.).

Jede Spalte verfügt über ihre eigene Suchfunktion, sodass sprachliche Einheiten im Textcorpus ihren Eigenschaften nach gesucht und gefiltert angezeigt werden können (z.B. über ihre Token-ID, ihre Wortart, ihr Grundwort, über morphologische Merkmale usw.). Das Zeichen „%“ kann während des Suchprozesses als Variable (Platzhalter) fungieren – und zwar für beliebig viele alphabetische

Zeichen. Es kann bei der Abfrage (auch gleichzeitig) sowohl am Anfang eines Wortes, im Wortinneren wie auch am Wortende verwendet werden, wobei es für beliebig viele beliebige Zeichen ab jener Position im Wort steht, an der es eingegeben wird. Es können, wie gesagt, auch mehrere Platzhalter „%“ gleichzeitig an verschiedenen Positionen des Wortes verwendet werden. Dies hat folgenden Vorteil: Besteht etwa in einem Textcorpus frühneuhochdeutscher Sprache die Wahrscheinlichkeit, dass ein und dasselbe Grundwort (Lemma) in mehreren voneinander abweichenden Schreibvarianten vorkommt, so hat man bei der Suche nach den betreffenden Tokens mittels Platzhalter die Möglichkeit, sich mit *einer* Suchabfrage *alle* unterschiedlichen Schreibvarianten anzeigen zu lassen. Suchanfragen erfolgen entweder manuell (beliebige Eingabe in das Suchfeld einer Spalte) oder anhand der vorgegebenen, verwendeten Tags in Drop-Down-Menüs (z.B. Auswahl des Wortarten-Tags „ADV“ für Adverbien > Anzeige aller mit „ADV“ getaggt Tokens). Mehrere Suchkriterien können miteinander kombiniert werden (z.B. Suche nach allen Adjektiven männlichen Geschlechtes, die im Akkusativ Plural stehen). Die Auswahl, die in den Drop-Down-Menüs verfügbar ist, ist dabei nicht fix vorgegeben, sondern hängt ganz davon ab, mit welchen Informationen (Tagsets) man den TokenEditor im Vorfeld der eigentlichen Annotation bespielt, also davon, wie man einen Text annotieren will – mit welchem Tags. Die Summe aller in den Drop-Down-Menüs sichtbaren Tags sollte in der Regel das vollständige Tagset ergeben, das verwendet wird.

Der TokenEditor ist demnach für verschiedene Arten tokenbasierter linguistischer Annotation offen.

Je mehr linguistische Merkmale man explizieren will, desto größer ist die Anzahl der benötigten Spalten. Dies wird bei der Arbeit an kleineren Bildschirmen mitunter unpraktisch, denn je mehr Spalten man verwendet, desto schmaler werden sie. Für die praktische Annotationsarbeit, besonders in Teams, ist es von größtem Vorteil, eine Spalte zu verwenden, die den Bearbeitungsstatus der einzelnen Tokens wiedergibt. Farbliche Unterlegungen haben sich in diesen Zusammenhang bewährt: In der Spalte „Status“ wird mittels Drop-Down-Menü vermerkt, in welchem Bearbeitungsstatus ein Token sich zum Zeitpunkt X befindet (z.B. „OK“ mit grüner Zeilenunterlegung oder „Questionable“ in Kombination mit grauer Farbe od. dgl. mehr).

Gegebenenfalls kann man in einer eigenen Spalte für Anmerkungen händisch eintragen und begründen, inwiefern ein Token noch nicht fertig bearbeitet bzw. zweifelhaft ist.

In den TokenEditor kann ferner die Möglichkeit implementiert werden, über einen Link in einer eigenen Spalte jene Digitalisate von Originalquellen anzeigen zu lassen, deren transkribierter Text in den TokenEditor eingespielt wurde. Diese Option ist vor allem dann sinnvoll, wenn der Verdacht besteht, dass während der Transkription Fehler passiert sein könnten und diese auf solche überprüft werden muss. Ferner gibt es die Möglichkeit, den TokenEditor mit anderen Tools, in die derselbe Text (einschließlich Annotationen!) eingespielt wurde, zu verlinken, z.B. mit der Sketch Engine, die vor allem für die Anzeige des sprachlichen Kontexts der Tokens von Vorteil ist (Listen im Stile einer Konkordanz).

Neben der Textdarstellung in vertikaler Form bietet der TokenEditor auch die Möglichkeit, über ein eigenes Feld „Context“ in der rechten oberen Ecke des Internet-Browsers den weiteren sprachlichen Kontext eines Tokens anzeigen zu lassen. Dieser Kontext ist für eine korrekte linguistische Annotation unerlässlich und von größter Wichtigkeit. Sobald man in der senkrechten Tabellendarstellung auf einen Token klickt, erscheint dieser grau unterlegt. Im Feld „Context“ rechts oben wird gleichzeitig der umgebende Text angezeigt, der mehr Tokens umfasst, als die senkrechte Darstellung auf einmal wiedergeben kann (wenn man den Scroll-Balken nicht bewegt). Der angeklickte Token erscheint in seinem sprachlichen Kontext grün unterlegt.

Zwischen den einzelnen Spalten des TokenEditors kann man während des Annotationsprozesses entweder mit den Pfeiltasten hin- und herwechseln, durch Maus-Klicks oder durch Druck der Tabulatortaste. Bei der manuellen Annotation kann man entweder dem Text folgen und jeden Token einzeln mit den entsprechenden linguistischen Informationen versehen (in der Reihenfolge des Auftretens) oder man lässt sich über die erwähnten Suchfunktionen alle Belegstellen für Tokens bzw. Lemmata anzeigen, die häufiger in einem Textcorpus vorkommen, z.B. Füllworte, bestimmte Adverbien, Einleitungsworte, Relativpronomina et c.; der TokenEditor zeigt an, wie oft jeder einzelne Token bzw. jedes Lemma in einem Textcorpus belegt ist und zeigt alle Belegstellen an, wenn man entsprechende Suchabfragen eingibt. Weiß man ungefähr, welche Tokens in einem

Textcorpus häufiger vorkommen (z.B. unter Zuhilfenahme der Sketch Engine), so lässt man sich zur schnelleren Bearbeitung alle Belegstellen für ein und dieselbe sprachliche Einheit gleichzeitig anzeigen. Über spezielle Eingabefelder in der rechten unteren Ecke des Internet-Browsers („Batch Edit“) kann man alle linguistischen Merkmale, die die angezeigten sprachlichen Einheiten gemeinsam haben, eingeben. Durch Anklicken der Schaltfläche „Assign to all“ werden diese *einmal* eingegebenen Merkmale *allen* markierten Tokens derselben Art zugeordnet. Man muss also nicht jeden Token einzeln händisch annotieren, wenn er mehrmals und immer in der gleichen Art und in der gleichen Funktion vorkommt, was sehr zeitaufwendig wäre, sondern man erspart sich durch die Annotation über „Batch Edit“ viel Arbeit. Diese Funktion erfordert es vor ihrer Benutzung jedoch, dass alle angezeigten und markierten Tokens sorgfältig überprüft werden, ob sie die in „Batch Edit“ manuell eingegebenen linguistischen Merkmale auch tatsächlich aufweisen. Bei ungenügender Prüfung der Tokens kann die erwähnte Funktion auch eine Fehlerquelle darstellen (flächige Falschzuweisungen).

Neben den beschriebenen Funktionen verfügt der TokenEditor auch über die Möglichkeit des Datei-Imports und des Exports (anwählbare Export-Formate: XML, CSV, JSON). Für die Benutzung dieser Funktionen bestand für den Verfasser vorliegender Projektbeschreibung jedoch keine Notwendigkeit.

Vorteilhaft bei der Verwendung des TokenEditor ist ein möglichst großer Bildschirm. Die einzelnen Spalten des Tools können bei einem möglichst großen Bildschirm alle in ausreichender Breite angezeigt werden (manuell mit der Computer-Maus einstellbar). Besonders bei der Explizierung morphologischer Merkmale ist eine beträchtliche Anzahl an Spalten notwendig. Für die Annotation der Eckhart-Briefe waren es deren insgesamt dreizehn (einschließlich Spalten für Token-ID, Token, Bearbeitungsstatus, Lemma und Wortart).

Die Kennzeichnung mehrteiliger, aus mehreren Tokens bestehender Eigennamen als zusammengehörige Signifikanten für eine (!) *named entity* erfordert darüber hinaus zusätzliche Spalten, in denen Beginn, Mitte und Ende der Zusammengehörigkeit markiert wird. Will man außerdem Spalten für Referenzwörterbücher oder für individuelle Anmerkungen zum Bearbeitungsstatus einzelner Tokens, Spalten für Verlinkungen mit digitalisierten Bildern oder mit anderen Tools implementieren, so kann dies auf Grund steigender Spalten-

Zahl in Verbindung mit morphologischer Annotation zu Problemen bei der Darstellung im Web-Browser führen, speziell bei kleinen Bildschirmen: Je mehr Spalten benötigt werden, desto schmaler werden sie angezeigt.

Möglichkeiten, linguistische Annotation mit anderen Annotationsschwerpunkten zu kombinieren, wie vom Verfasser der vorliegenden Projektbeschreibung ursprünglich intendiert, sind im TokenEditor zwar prinzipiell gegeben, aber aus den genannten praktischen und technischen Gründen begrenzt. Dies entspricht ganz dem Grundgedanken der Konzeption des Tools.

Für linguistische Basisannotation mit Kennzeichnung der Wortarten und mit Lemmatisierung sowie für morphologische Annotation stellt der TokenEditor ein Werkzeug erster Wahl dar. Abgesehen von diesen Annotationsarten sollten andere Annotationen aber direkt in Oxygen gemäß (TEI-)XML vorgenommen werden.

Für linguistische Basisannotation mit Kennzeichnung der Wortarten und mit Lemmatisierung sowie für morphologische Annotationen ist der TokenEditor in Summe sehr gut geeignet. Diese Annotationsarten zählen zu seinen Hauptanwendungsgebieten. Die erwähnten Möglichkeiten, Verlinkungen zu digitalisierten Bildern sowie zu anderen Tools zu implementieren, können dazu beitragen, den Annotationsprozess zu optimieren. Im Rahmen der linguistischen Annotation erweist der TokenEditor sich in Summe als ein äußerst leicht zu bedienendes und flexibles Tool, das individuell auf die Bedürfnisse der Annotatorinnen und Annotatoren abgestimmt werden kann.

2.4) Zur Annotation der Eckhart-Briefe

Die automatisch generierte linguistische Annotation der um editorische Zusätze bereinigten und automatisch tokenisierten Briefe Johann Georg Eckharts an Bernhard und Hieronymus Pez OSB erfolgte zunächst gemäß der Parameter des *nicht* erweiterten Frankfurter Tagsets sowie unter Zuhilfenahme des Münchner Tagging-Tools MarMot²¹⁶ (auf Veranlassung des ACDH). Die Ergebnisse dieses Annotationsprozesses sind über die Benutzeroberfläche des TokenEditors sichtbar und korrigierbar (nach passwortgeschütztem Log-in).

Sofern weitere bzw. differenziertere Tags für die Annotation nötig waren, als im ursprünglichen Tagset vorgesehen, wurden diese im TokenEditor ergänzt. Sie

²¹⁶ Vgl. Anm. 16.

wurden auf dessen Veranlassung von Mitarbeitern des ACDH in den TokenEditor implementiert. Anhand des erweiterten Frankfurter Tagsets wurden die Ergebnisse des automatischen Annotationsprozesses händisch korrigiert und ergänzt. Es erfolgten Lemmatisierung, Wortarten-Tagging und die Explizierung morphologischer Merkmale.

2.4.1) Einzelprobleme und Schlussfolgerungen

Wie bereits angedeutet bringt die Annotation historischer Textcorpora im Laufe des Annotationsprozesses spezielle Herausforderungen mit sich.²¹⁷ In welcher Weise – dies hängt davon ab, ob und inwiefern ein Text sich an einem (wie auch immer gearteten) sprachlichen Standard orientiert bzw. inwiefern er Nonstandard-varietäten aufweist. Corpora moderner Standardsprachen stellen für Tagger jüngerer Generation in der Regel kein Problem mehr dar. Ihre Trefferquoten liegen bei automatischer Annotation häufig im Bereich von über 90%. Mit leichten Vorbehalten gilt dies im Normalfall auch für lateinische Corpora, wenn deren Sprache sich weitestgehend an den Regeln der „klassischen“ und durchwegs weit verbreiteten Norm-grammatik orientiert.

Die beträchtliche Anzahl an vorhandenen Tagsets für die lateinische Sprache sowie der darauf trainierten Tagger verweisen zudem darauf, dass klassische Philologinnen und Philologen sich teilweise schon recht früh mit Natural Language Processing befasst haben. Gleichwohl konnten sich bisher kein einheitlicher Standard und kein einheitliches Tagset für die Annotation lateinischer Texte durchsetzen.²¹⁸ Die Texte neulateinischer Gelehrtenbriefe orientieren sich idealiter an einem sehr hohen Standard einer als klassisch empfundenen Latinität, vor allem in Hinblick auf Morphologie und auf Orthographie. Einflüsse frühneuzeitlicher Vernakularsprachen in der neulateinischen Syntax fallen bei der tokenbasierten Annotation punktuell zwar auf, können in deren Rahmen aber (vorläufig) vernachlässigt werden. Dasselbe gilt für

²¹⁷ Vgl. auch *Czeitschner, Resch*, Austrian Baroque Corpus (wie Anm. 9), 40f.

²¹⁸ Bemühungen zur Vereinheitlichung der sehr zahlreichen und vielfältigen, aber ebenso dispersen und uneinheitlichen lateinischen DH-Projekte der letzten Jahre und Jahrzehnte werden derzeit von einer Forscher*innengruppe rund um Marco Passarotti und das von ihm angeführte Projekt „LiLA – Linking Latin“ unternommen (siehe unten).

syntaktische Phänomene, die – bei allem Anspruch der Briefschreiber – punktuell nicht immer den als klassisch empfundenen Normen entsprechen, jedoch nicht auf den Einfluss von Vernakularsprachen zurückzuführen sind.

Wollte man hinsichtlich der Lexik neulateinischer Gelehrtensprache aus der Frühen Neuzeit ausschließlich auf die gängigsten Lateinwörterbücher referenzieren, z.B. ausschließlich auf den Georges²¹⁹, der sich hauptsächlich am klassischen Latein orientiert, so stellt sich bald die Erkenntnis ein, dass solche Wörterbücher dafür nicht ausreichen. Weitere müssen hinzugezogen werden, etwa der DuCange²²⁰ sowie mittellateinische Wörterbücher. In manchen Fällen wird man sich damit begnügen müssen, dass einzelne Lemmata in keinem verfügbaren Wörterbuch nachweisbar sind („*out of vocabulary*“).

Die Annotation von Neulatein gestaltet sich im Vergleich zur Arbeit mit frühneu-hochdeutschen Texten insofern einfacher, als im ersten Fall die Konventionen der Zusammen- und der Getrennschreibung kaum Schwankungen unterworfen sind und einen hohen Grad an Standardisierung aufweisen. Generell ist bei lateinischen Texten allerdings zu beachten, dass die Enklitika *-que/-ne/-ve* u. dgl. als eigene Tokens zu führen sind.

Orthographische Abweichungen von der klassischen Norm bleiben in den Eckhart-Briefen in der Regel auf einzelne Konsonanten, Vokale oder Diphtonge beschränkt. Zumeist finden sie sich in Substantiven. Diese Abweichungen können teilweise auf Gepflogenheiten der spät- sowie der mittellateinischen Gelehrtensprache zurückzuführen sein, etwa der Kirchenväter, so zum Beispiel hyper-urbanes *ae* statt *e* in *caeteri*, dagegen: *e* statt *ae* in *seculum*, *o* statt *u* in *epistola*, *gracia* statt *gratia*; *author* statt *auctor* u. dgl. mehr. Worte mit derartigen Abweichungen stellten den Tagger MarMot vor keinerlei Probleme und konnten einer automatischen Lemmatisierung ohne weitere Schwierigkeiten unterzogen werden. Dies gilt vor allem für Substantive, Adjektive, Adverbien und Pronomina.

²¹⁹ Zuletzt als: Der Neue Georges. Ausführliches lateinisch-deutsches Handwörterbuch (2 Bde.), ursprüngl. ausgearb. v. Karl-Ernst Georges, bearb. v. Heinrich Georges, neu hrsg. v. Thomas Baier & bearb. v. Tobias Dänzer (Darmstadt 2013).

²²⁰ Z.B. in folgender Ausgabe: Charles Du Fresne Du Cange, Lorenz Diefenbach, *Lexicon mediae et infimae latinitatis* (Darmstadt 1997; unveränd. reprogr. Nachdr. d. Ausg. Frankfurt a.M. 1857).

Anders verhielt es sich zuweilen im Bereich der Verben. Dort kam es während der automatischen Lemmatisierung häufiger zu Fehlern. Bemerkenswert ist in diesem Zusammenhang, dass auch falsch lemmatisierte Verben in den allermeisten Fällen mit korrekter Morphologie versehen wurden. Morphologische Fehlzusweisungen traten zumeist im Bereich zweideutiger Casusendungen auf (Liste ausgewählter maschineller Fehler im Anhang).

Probleme seitens der automatischen Annotation traten erwartungsgemäß auch dort auf, wo Johann Georg Eckhart in seinen Briefen zwischen lateinischer und deutscher Sprache wechselte: Stellenweise wurden diese Passagen von MarMot nicht nur nicht oder nicht nur falsch annotiert, sondern überhaupt ausgelassen, also nicht als Bestandteil des Textes erfasst und überhaupt nicht tokenisiert, sodass sie im TokenEditor zunächst gar nicht aufschienen, was eine weitere Vorverarbeitung nötig machte – und zwar insofern, als die deutschen Passagen im Falle eines Briefes ausgeklammert werden mussten, auf dass zumindest die lateinischen digital verarbeitet werden konnten (Brief „GE 10“ de dato 1719-01-12 – vgl. Anm. 15).

Im Falle mehrsprachiger Corpora ist demnach in erhöhtem Maße auf exakten Abgleich mit den Quellen zu achten. Je nach Sprache kann auf andere digitale Tools zurückgegriffen werden. Die im Folgenden angestellten Berechnungen zur Treffer- bzw. zur Fehlerquote von MarMot bei einzelnen sprachlichen Kategorien (Casus, Numerus et c.) beziehen sich demnach ausschließlich auf *lateinische* Passagen in Eckharts Briefen.

Im Anschluss an Claudia Resch und Wulf Oesterreicher kann der Verfasser der vorliegenden Zeilen auf Grund seiner Annotationspraxis mit den Eckhart-Briefen bestätigen, dass es sich bei der Annotation historischer Sprachdaten um eine sehr „komplexe, höchst anspruchsvolle Forschungsaufgabe“²²¹ handelt. In der Theorie wird die Vielschichtigkeit dieser Aufgabe zuweilen unterschätzt. Denn neben soliden sprachhistorischen Kenntnissen, historisch-kulturellem Wissen, Vertrautheit mit den Texten und viel Feingefühl bei der Beurteilung einzelner, sprachlicher Phänomene sind vielfach auch pragmatische Lösungen gefragt. Diese pragmatischen Lösungen sind zum Teil einer einfachen Handhabung der elektronischen Werkzeuge geschuldet, mit denen gearbeitet wird.

²²¹ Czeitschner, Resch, Austrian Baroque Corpus (wie Anm. 9), 49f.

Zu umfangreiche Tagsets erschweren die automatische Annotation und verlängern jene Zeit, die benötigt wird, um die automatischen Tagging-Ergebnisse zu generieren, wovon gleichzeitig deren Korrektheit beeinträchtigt werden kann. Pragmatische Lösungen sind ebenso der Konsistenz jener Annotationsgrundsätze geschuldet, die man sich selbst als Richtlinie bei der Annotation setzt.

Denn entgegen der eigenen, ursprünglichen Annahme ist ein Höchstmaß an feingranularer, linguistischer Pointiertheit, Treffergenauigkeit und Differenziertheit offenbar nicht immer bis in alle Einzelheiten erstrebens- bzw. wünschenswert. Dies liegt wiederum an der Eigenart der Materie „Sprache“ selbst. Dem Gebrauch ihrer Bestandteile, den Worten im weitesten Sinne, sind Ambiguitäten naturgemäß in höchstem Maße inhärent. Versteht man Digitalisierung als die Umwandlung komplexer Phänomene in eindeutig distinkte, eindeutig abgrenzbare und eindeutig zählbare Einheiten, so muss man offensichtlich eingestehen, dass Sprache sich in mancherlei Hinsicht einer strengen Digitalisierung entzieht – und zwar insofern, als es beispielsweise nicht immer leicht fällt, Zeichenketten *einer* (!) bestimmten Wortart exakt zuzuordnen.

Welche Wortart man einer bestimmten Zeichenkette in Einzelfällen konkret zuweist, hängt abgesehen vom sprachlichen Kontext und ihrem punktuellen Gebrauch auch davon ab, welcher klassischen Grammatik man zu folgen geneigt ist, mit welchen Ansprüchen man linguistisch annotiert – und nicht zuletzt auch von der Entscheidung des Annotators/der Annotatorin.

Im Lehrbuch der lateinischen Syntax und Semantik von Hermann Menge²²² scheinen bestimmte Zahlworte wie *unus, primus, singuli* u. dgl. als Untergruppe der Adjektive auf, ebenso unbestimmte Zahlworte, die entweder substantivisch oder adjektivisch verwendet werden können (wie *omnes, cuncti, alii, multi, ceteri, reliqui, pauci, plures, complures, nonnulli* u. dgl.). Beim Tagging stellt sich in der Praxis die Frage, mit welchen Tags solche Worte zu versehen sind.

Ein Tag ADJ kommt bei hauptwörtlichem Gebrauch streng genommen nicht in Frage. Ein Tag wie etwa ADJ_NN ist gegebenenfalls zu verwenden, kommt aber in keinem der evaluierten Tagsets vor und musste im Rahmen des hier durchgeführten Projektes ergänzt werden. Wollte man dagegen betonen, dass es

²²² Z.B. in folgender Ausgabe: Hermann Menge, Lehrbuch der lateinischen Syntax und Semantik, bearb. v. Thorsten Burkhard, Markus Schauer (5. durchges. u. verb. Aufl. Darmstadt 2012).

sich um Adjektive handelt, die für eine (nicht) näher definierte Zahl bzw. Menge stehen, so könnte man beispielsweise NUM (numeral) verwenden.

Allerdings müsste dann näher zwischen definiten und indefiniten Zahladjektiven unterschieden werden, bei den bestimmten ferner zwischen Kardinal-, Ordinal-, Distributiv- und Multiplikativzahlworten. Eine linguistisch gesehen höchst feingranulare Annotation ist also mit einer ganz beträchtlichen Anzahl an Tags verbunden.

Auch dann, wenn es gilt, hauptwörtlich gebrauchten Perfektpartizipien mit eigenständiger Bedeutung einen Wortarten-Tag zuzuweisen, sind pragmatische Entscheidungen seitens der Annotatorinnen und Annotatoren gefragt. Bei Zeichenketten wie etwa *acta* (Akten), *excerpta* (Exzerpte), *coniunctus* (Verwandter) oder *praepositus* (Vorsteher) handelt es sich der Wortwurzel nach um Verben (Tag: V). Gleichwohl können solche Zeichenketten ihrem Gebrauch nach als Substantive getaggt werden (z.B. mit NN – normal noun). Der Aspekt, dass es sich im Falle des Grundworts um ein Verb handelt, kann über das morphologische Tagging zum Ausdruck gebracht werden (z.B. Annotation PART für Partizip).

Bei der Lemmatisierung ist ferner zu entscheiden, ob man die partizipiale Form als Grundwort beibehält – *acta* würde demnach mit *acta* lemmatisiert – oder ob man das Grundwort über etymologische Kriterien vergibt (z.B. *ago* bzw. *agere* als Lemma für *acta*?). Aus Phänomenen wie diesen wird klar ersichtlich, dass das Tagging lateinischer Texte zu einer fundierten Auseinandersetzung mit mehreren Faktoren zwingt: erstens mit dem konkreten Gebrauch einer Zeichenkette im Text der Quelle; zweitens mit den linguistischen Gegebenheiten der verwendeten Sprache; drittens mit den Kategorien traditioneller Grammatiken, die nicht immer frei von arbiträren Entscheidungen und auch nicht immer frei von Widersprüchen im Vergleich zueinander sind; viertens mit Fragen nach linguistisch feingranularer Annotation gegenüber einer pragmatischen, aber verkürzenden Zugangsweise, die der zwar Schnelligkeit und der Genauigkeit der *automatischen* Datenverarbeitung entgegenkommt, linguistisch gesehen aber oberflächlicher ausfallen muss.

Die fundierte Abwägung dieser und anderer Faktoren (am besten unter mehreren fachlich geschulten Wissenschaftlerinnen und Wissenschaftern) muss letztlich zu einer begründeten und nachvollziehbaren Entscheidung für die Erstellung von

Tagsets und von Annotationsrichtlinien führen. Diese Richtlinien sind in weiterer Folge möglichst konsequent und einheitlich umzusetzen. Zweifelsfälle sind dabei im Team immer wieder zu besprechen und abzugleichen.

In diesem Zusammenhang ist allerdings einzuräumen, dass bei der Erstellung eines Tagsets *vor* der Annotation eines Textes *alle* sprachlichen Phänomene nur in den allerseltensten Fällen a priori lückenlos werden erfasst werden können. Stärker noch als die Annotation von neulateinischen Texten hat der Umgang mit älteren Sprachstufen des Deutschen dem Verfasser vorliegender Zeilen gezeigt, dass es immer wieder von Nöten sein kann, Tagsets *während* des Arbeitsprozesses um zusätzliche Tags zu erweitern, vor allem dann, wenn die behandelte Sprachstufe einen verhältnismäßig geringen Grad an Standardisierung aufweist und/oder der Urheber/die Urheberin eines Textes ein eigenständiges und kreatives Moment einbringt, beispielsweise bei belletristischen Texten oder insbesondere bei Lyrik.

Inwiefern eine linguistisch feingranulare Annotation einschließlich Syntax und Morphologie für editionswissenschaftliche Zwecke zum Einsatz kommt, wird in der Praxis von mehreren Faktoren abhängen, nicht zuletzt auch von personellen, zeitlichen und finanziellen Ressourcen, andererseits von den Zielsetzungen der Edition, die kein Referenzcorpus sein kann bzw. sein möchte, sondern allenfalls ein historisches Spezialcorpus.

Für linguistisch-komparative Fragestellungen mittels digitaler Suchabfragen bieten morphologisch annotierte Parallelcorpora einen vielversprechenden Ausgangspunkt. Allerdings ist stets der Aufwand zu bedenken, der mit morphologischer Annotation einhergeht. Gleichzeitig muss man sich, wie oben skizziert, die Begrenztheit vor Augen halten, die trotz morphologischer token-basierter Annotation hinsichtlich semantischer Aussagekraft gegeben ist.

Syntaktisch annotierte Treebanks sind, wie weiter oben ebenfalls bereits erwähnt, für die Editionswissenschaft wegen des beträchtlichen Aufwandes bei ihrer Erstellung so gut wie impraktikabel. Treebanking ist allenfalls bei nachträglicher Digitalisierung von bereits edierten Texten mit einigermaßen rezenten Textausgaben als Ausgangsmaterial anzuraten, weil in diesem Fall der Aufwand des editorischen Kerngeschäftes minimiert wird bzw. ganz wegfällt, nämlich das Transkribieren und das Kollationieren.

Indes haben Projekte wie ABaC:us (vgl. Anm. 9 und 92) gezeigt, dass linguistische Basisannotation mit Lemmatisierung und mit Auszeichnung der Wortarten für die Unterstützung und/oder für die Ergänzung bisheriger Forschungsergebnisse linguistischer, historischer und/oder literaturwissenschaftlicher Natur durchaus ausreichend sein kann.

2.4.2) Maschinelle Annotation, händische Eingriffe und Änderungen

Im Hintergrund der graphischen Benutzeroberfläche des TokenEditor werden jene Änderungen protokolliert und gespeichert, die man nach der automatischen Vorannotation händisch an den linguistischen Merkmalen (Attributen) der einzelnen Tokens vornimmt. Diese Änderungen lassen sich ihrerseits als XML-Datei darstellen und geben den manuellen Annotationsverlauf detailliert wieder. Auf ihrer Basis lassen sich (mit gewissen Vorbehalten) Berechnungen zur Trefferquote automatischer Tagging-Tools anstellen. Im Falle der annotierten Eckhart-Briefe handelte es sich, wie gesagt, um den Münchner Tagger MarMot. Die verwendeten Daten können der Forschung auf Anfrage bei Bedarf zur Verfügung gestellt werden.

Bei der Bewertung der händisch korrigierten linguistischen Attribute ist insofern Vorsicht angebracht, als viele manuelle Änderungen deshalb vorgenommen wurden, weil der Verfasser der vorliegenden Zeilen mit einer erweiterten Version des Frankfurter Tagsets operierte, auf die MarMot bisher nicht trainiert ist. Tags aus dieser erweiterten Version konnten, wie gesagt, nicht automatisch, sondern ausschließlich manuell zugeordnet werden.

So ergibt es sich, dass das am häufigsten geänderte linguistische Attribut die Wortart (PoS = type) ist (2.020 Änderungen). Der Bedarf, diese Änderungen vorzunehmen, ist jedoch in vielen Fällen dem erweiterten Tagset geschuldet und darf bei einer Berechnung der Fehlerquote von MarMot insofern nicht mit einbezogen werden.

Die zweithäufigsten Änderungen betrafen den Bereich der Grundworte (Lemmata), die den einzelnen Tokens zugeordnet wurden (1.389 Änderungen). Auch hier ist es jedoch nicht angebracht, ausgehend von allfällig vorgenommenen händischen Korrekturen den Schluss zu ziehen, während der automatischen

Annotation seien seitens des Taggers Fehler passiert. Die begründete und einheitliche Wahl der Grundworte stellt eine äußerst komplexe Aufgabe dar. Eine Entscheidung hat letztlich seitens des Annotators/der Annotatorin zu erfolgen, so etwa im Falle von Adjektiven mit unregelmäßiger Steigerung. Insofern stellt es keinen maschinellen Fehler dar, wenn entschieden wird, eine flektierte Form von *plus* nicht mit *plus*, sondern mit *multus* zu lemmatisieren und *melior* im Lemma auf *bonus* zurückzuführen.

Sehr deutlich lassen maschinelle Fehler sich an den Kategorien Gender/Geschlecht, Case/Casus, Tense/Zeit und Numerus/Zahl ablesen, da diese Kategorien nicht der Entscheidung des Annotators bzw. der Annotatorin unterliegen, sondern durch Sprache und Text vorgegeben sind. An insgesamt 7.913 vorhandenen, deklinierbaren Tokens (Substantiven, Adjektiven, Pronomina, Partizipien) in den Eckhart-Briefen wurden 880 Änderungen im Bereich der Casus vorgenommen, wenn diese von MarMot entweder nicht automatisch vorgetaggt wurden oder falsch. Das ergibt eine maschinelle Fehlerquote von rund 11,12%. Anders gesagt wurden rund 88,88% aller Casus von MarMot richtig annotiert. Gleichzeitig wurden in der Kategorie Gender 975 manuelle Änderungen vorgenommen. Das Geschlecht deklinierbarer Worte wurde von MarMot demnach in rund 85,74% aller Fälle richtig zugeordnet (Fehlerquote von rund 14,26%).

Bei Verben wurde die Kategorie Tense 252 Mal korrigiert. Bei einem Gesamtaufkommen von 2.531 Verbformen (flektierte und unflektierte zusammen) in den Briefen Johann Georg Eckharts bedeutet dies eine maschinelle Trefferquote von rund 90,04% (Fehlerquote: rund 9,96%).

In der Kategorie Voice (aktiv/passiv) wurden 241 manuelle Änderungen vorgenommen. Da aber zusätzlich zu den Tags des Frankfurter Tagsets ein eigener Tag für Deponentia ergänzt wurde (DEP), ist diese Zahl nicht repräsentativ.

Bei insgesamt 8.666 Tokens mit der Kategorie Numerus wurde die Zahl 664 Mal ergänzt bzw. korrigiert. Daraus ergibt sich für die automatische Annotation eine Trefferquote von rund 92,34% in dieser Kategorie (Fehlerquote: rund 7,66%).

364 Korrekturen bzw. Ergänzungen mussten in der Kategorie Mood vorgenommen werden. Da dieser Kategorie jedoch zusätzliche Tags eingefügt worden waren (INF-ACI und INF-NCI), wird hier keine automatische Fehler-/Trefferquote berechnet.

Unter Ausklammerung der Kategorien Lemma und (PoS-)Type ergibt sich anhand der hier errechneten Zahlen eine durchschnittliche Trefferquote von 89,25% für die automatische Vorannotation.

2.5) Corpusabfragen und -auswertung

Um auf Basis der linguistischen Annotationsergebnisse Aussagen quantitativer Natur zur Sprache tätigen zu können, die Johann Georg Eckhart in den lateinischen (!) Passagen seiner Briefe an Bernhard und Hieronymus Pez OSB verwendete, wurde vorweg eine Reihe von Forschungsfragen definiert. Diese lassen sich in Form computergestützter Corpusabfragen mittels TokenEditor beziehungsweise mittels Sketch Engine innerhalb der annotierten Briefe umsetzen. Morphologisch annotiert wurden rund 14.850 Tokens (mit dem erweiterten Frankfurter Tagset). Für linguistische Vergleiche wurden als Parallelcorpus jene Briefe in den TokenEditor eingespielt, die Bernhard und Hieronymus Pez OSB zwischen 1709 und 1715 an verschiedene Adressaten versandten. Dieses Corpus umfasst 23.414 Tokens. Es wurde ausschließlich maschinell annotiert. Eine differenzierte Annotation mit dem erweiterten Frankfurter Tagset konnte hier bis dato noch nicht durchgeführt werden.

Die Möglichkeit, beide Corpora im Detail, sprich: über morphologische Eigenschaften ihrer Bestandteile, miteinander zu vergleichen, ist vorerst also noch nicht vollständig gegeben. Über die Suchfunktion im TokenEditor lässt sich jedoch schon jetzt die Anzahl an Belegstellen bestimmter Zeichenketten im Pez-Corpus generieren. Wo dies möglich war, wurden im Rahmen der erwähnten Forschungsfragen (Details siehe unten) daher bereits Vergleiche zwischen Pez' und Eckharts sprachlichen Vorlieben angestellt, zumindest auf Basis der Suche nach Tokens und nach Lemmata. Dass das Pez-Corpus noch nicht vollständig annotiert wurde, ist dem Umfang des Gesamtprojektes geschuldet. Es soll mit den folgenden Forschungsergebnissen nur eine grundsätzliche Richtung angezeigt werden, welche man kraft der Verbindung von Editionswissenschaft und Corpuslinguistik einschlagen *kann*.

2.5.1) Forschungsfragen und quantitative Ergebnisse

Gefolgt von den dazugehörigen Antworten und dem prozentuellen Anteil bestimmter Tokens an der Gesamtzahl der annotierten sprachlichen Einheiten lauten die erwähnten Forschungsfragen folgendermaßen:

- Wenn Pez bzw. Eckhart explizit von ihren Quellen oder von ihren editorischen Werken sprechen, wie oft tun sie das, indem sie Begriffe wie *codex*, *liber*, *libellus* oder *manuscriptum* für die Überlieferungsträger bzw. für die edierten Werke verwenden? Wie oft tritt *volumen* in ihren Briefen auf, wie oft *diploma* oder *chronicon*? Welche der genannten Zeichenketten überwiegt quantitativ bei wem der beiden?
- *codex*: 30 Belege bei Eckhart (0,2%), 23 bei Pez (0,1%), davon fünf Belegstellen in einem einzigen Brief (de dato 1714-09-20, Bernhard Pez an Johann Christoph Bartenstein, in der Edition N. 360) und mit Bezug auf *einen konkreten Codex* aus dem Bestand der Wiener Hofbibliothek;
- *liber*: 14 Belege bei Eckhart (0,09%), 27 bei Pez (0,12%)
- *libellus*: 7 Belege bei Eckhart (0,05%), 7 Belege bei Pez (0,03%; zzgl. einer Belegstelle für *libellulus*, ein „doppelter“ Diminutiv)
- Wortstamm **manuscript-* : 10 Belege bei Eckhart für das Hauptwort (NN) *manuscriptum* (0,07%); als Adjektiv (ADJ) in Verbindung mit einem Hauptwort (z.B. *codex*) verwendet Eckhart die Zeichenkette *manuscriptus* überhaupt nicht; bei Pez tritt das Substantiv (NN) *manuscriptum* an sieben Belegstellen auf (0,03%); viel häufiger, nämlich 26 Mal (0,11%), verwendet Pez dagegen das Adjektiv *manuscriptus* (ADJ), beispielsweise in Verbindung mit Substantiven wie *chronicon* oder mit *codex*. Da der Wortstamm **manuscript-* eine Homographie darstellt (gleich bei ADJ und NN) und auch die lateinischen Flexionsendungen ohne Blick auf den sprachlichen Kontext an sich keinen Aufschluss darüber erlauben, ob mit der verwendeten Zeichenkette an einer Belegstelle ein Adjektiv oder ein Substantiv gebildet wird, zeigt sich der Nutzen insbesondere der Wortarten-Annotation und der Lemmatisierung mit digitalen

Werkzeugen, die hier zur Distinktion fähig sind. Zwar bedurften die automatisch generierten PoS-Tags im Falle des angeführten Beispiels stellenweise der händischen Korrektur. Doch grundsätzlich zeigt sich hier ein Vorteil von Taggern und Tagging-Ergebnissen gegenüber einfachen Konkordanz-Tools wie AntConc, denn bei Verwendung von Letzterem wäre der Unterschied zwischen Hauptwort und Adjektiv viel weniger deutlich zu Tage getreten.

- *volumen*: 6 Belege bei Eckhart (0,04%), 5 bei Pez (0,02%).
- *diploma*: 32 Belege (0,22%) bei Eckhart; zzgl. ein Beleg für das Adjektiv *diplomaticus*; bei Pez: 11 Belege (0,047%) für das Hauptwort; das Adjektiv *diplomaticus* tritt im untersuchten Pez-Corpus überraschenderweise nicht auf.
- *chronicon*: 8 Belege bei Eckhart (0,05%), 16 bei Pez (0,07%).
- Wie ist es um das Verhältnis zwischen unterordnenden und beiordnenden Bindeworten in den untersuchten Briefen Eckharts bestellt (Subjunktionen gegenüber Konjunktionen)?
 - Anzahl Subjunktionen (CS): 312 (2,1%) zzgl. 238 (1,6%) Relativpronomina (PREL);
 - Anzahl Konjunktionen (CC): 669 (4,51%).

Um auf Basis dieser Annotationsergebnisse Aussagen zum zahlenmäßigen Verhältnis von Haupt- und von Nebensätzen treffen zu können, müsste näherhin differenziert werden, wie viele beiordnende Bindeworte in Nebensätzen vorkommen.

Ferner wurde gefragt:

- Wie (oft) verwendet Eckhart in den untersuchten Briefen das Wort *cum* (bevorzugt – als Präposition oder als Einleitungswort)?
 - insgesamt 69 Mal (0,46%), davon 52 Mal als Präposition, 17 Mal als Einleitungswort.
 - Pez: insgesamt 83 Mal (0,35%).
- Wie oft verwendet Eckhart demgegenüber das Wort *ut*?
 - insgesamt 94 Mal (0,63%), davon 93 Mal als Subjunktion, einmal als

Konjunktion im Hauptsatz (= „wie“)

- Pez: 190 Mal (0,81%).
- Wie oft kommt das Wort *quod* als Einleitungswort bei Eckhart vor?
 - Bei 43 Belegen (0,29%) für *quod* insgesamt 13 Mal als Subjunktion (CS) und 28 Mal als Relativpronomen (PREL), zwei Mal als Demonstrativpronomen.
 - Pez: insgesamt 141 Mal (0,60%).

2.5.2) Semantische Einzelanalysen

Als Beispiele für eingehendere corpusbasierte Untersuchungen sollen an dieser Stelle die Semantik sowie der sprachliche Kontext analysiert werden, in welchem Johann Georg Eckhart und Bernhard Pez die Begriffe *codex* und *liber* (einschließlich abgewandelter Formen und einschließlich Diminutiva) verwendeten – sowie Begriffe, die für eine Assoziation mit den genannten Termini in Frage kommen: *charta/chartaceus*, *membrana/membranaceus* und *folium*.

Diese Forschungsfragen könnten zwar auch gänzlich ohne Zuhilfenahme digitaler Methoden und Werkzeuge beantwortet werden. Ein linguistisch annotiertes Corpus ermöglicht es jedoch, die entsprechenden Belegstellen für gesuchte Worte um ein Vielfaches schneller ausfindig zu machen als mit herkömmlichen Methoden. Dabei zeigt sich vor allem die zentrale Bedeutung der Lemmatisierung: Durch sie kann durch Eingabe eines nicht flektierten Grundwortes, des Lemmas, jede abgewandelte Form gleichzeitig angezeigt werden.

Allerdings ist im vorliegenden Zusammenhang mit Nachdruck darauf hinzuweisen, dass mit den hier genannten Stichworten längst nicht alle sprachlichen Möglichkeiten umfasst sind, derer sich Pez und Eckhart in ihren Briefen bedienen konnten, um über die Quellen ihrer historischen Forschungen zu sprechen. Die Frage etwa, inwiefern Überlieferungsträger *umschrieben* werden, indem beispielsweise ein in ihnen enthaltenes *Werk* durch Angabe seines Titels genannt wird, wird im Rahmen der vorliegenden Studie vorerst nicht behandelt, da dafür umfangreichere Annotationen (nicht nur linguistischer Natur) erforderlich wären.

Als *sprachlicher Kontext* im obigen Sinne sei hier die linguistische Umgebung im Umfang von jenen ein bis zwei Sätzen definiert, die einen Satz mit den gesuchten Termini beiderseits umgeben (vorher und nachher).

Zu Eckharts Briefen an Pez lässt sich vorweg zusammenfassend sagen, dass sich in der Verwendung des Begriffes *codex* ein breites Spektrum an Tätigkeiten und Interessen manifestiert, wie sie für einen philologisch, historisch-kritisch und editorisch arbeitenden Gelehrten des 18. Jahrhunderts typisch waren.

Insgesamt lässt sich beobachten, dass Eckhart häufig auf *ganz bestimmte, konkrete* Überlieferungsträger referenziert, wenn er den Begriff *codex* verwendet: Von 30 Belegen stehen 23 im Singular, sieben im Plural. Eckhart bezeichnet damit fast ausschließlich handschriftliche Überlieferungsträger, nur einmal wird auf eine Edition referenziert (*Aliqua Leibnitius in Codice diplomatico profert* – Eckhart an Hieronymus Pez, 1718-12-25, Hannover; in der Edition: Brief 1031 = „BE 11“).²²³

Der Semantik nach scheint *manuscriptum* daher nicht unähnlich, doch lässt sich im Falle dieses Terminus ohne eingehendere Studien, die auch die Editionen Eckharts umfassen müssten, nicht klären, worin Unterschiede bestehen könnten. Es ist nicht auszuschließen, dass *manuscriptum* nicht nur gebundene Codices, sondern insbesondere auch Teile daraus, Fragmente oder ungebundene Textträger bezeichnet. Dies soll in vorliegender Studie aus Platzgründen vorerst aber unberücksichtigt bleiben.

Analysiert man den Kontext der Belegstellen für *codex* semantisch genauer, so lässt sich feststellen, dass Angaben zum Aufbewahrungsort, zur Herkunft und/oder zur Besitzgeschichte der angesprochenen Codices am häufigsten auftreten, gefolgt von Angaben zu deren Inhalt.

²²³ Eine Liste ausgewählter Belegstellen, auf die in den folgenden Ausführungen Bezug genommen wird, findet sich in den Anhängen zu vorliegender Arbeit. Um allzu viele Einschübe bzw. Fußnoten zu vermeiden, wird bei Bezugnahme auf Originalstellen aus der Pez-Korrespondenz auf die Liste dieser Belegstellen im Anhang verwiesen. Die Siglen für diese Belegstellen lauten „BE [Ziffer]“ bzw. „BP [Ziffer]“. „BE“ steht für „Belegstelle Eckhart“, „BP“ für „Belegstelle Pez“. Die Nummerierung der Belegstellen folgt der Logik und der Reihenfolge ihrer Bearbeitung in vorliegender Arbeit – auf Basis der elektronischen Daten, nicht auf Basis der Chronologie der Briefe in der Edition. Die Liste der Belegstellen im Anhang verfügt über Hinweise zur Auffindung jener Briefe, in denen die Belegstellen enthalten sind: Datum und Ort der Abfassung, Absender, Empfänger sowie Nummerierung der Briefe in den gedruckten Editionen. Warum die Reihenfolge der Briefe in den elektronischen Daten der Chronologie in der Edition teils nicht entspricht, konnte in vorliegender Arbeit aus organisatorischen Gründen nicht mehr geklärt werden. Ein Grund ist jedenfalls in der elektronischen Vorverarbeitung der frühen Pez-Briefe zu suchen, die nicht vom Verfasser vorliegender Arbeit vorgenommen wurde.

Dass Angaben zur Herkunft und/oder zum Aufbewahrungsort so zahlreich vertreten sind, ist darauf zurückzuführen, dass Eckhart im Rahmen seiner gelehrten Reisen immer wieder Stätten seines Interesses aufsuchte, so etwa Klöster, um sich über Bestände in deren Bibliotheken und/oder in deren Archiven zu informieren und sich diese bei entsprechender Gelegenheit auch zeigen zu lassen, wenn möglich entweder als ganzes oder wenigstens in Form von ausgesuchten Einzelstücken. Die auf diesem Wege gewonnenen Einsichten gibt er brieflich häufig an Bernhard Pez weiter.

Dabei macht Eckhart nur selten präzise Angaben zum Alter der in Augenschein genommenen Codices (drei Mal mit Angabe des Jahrhunderts: BE 2; BE 29; BE 31). Ein Grund hierfür ist darin zu sehen, dass Eckhart während seiner Reisen zu wenig Zeit aufwenden konnte, um die Überlieferungsträger eingehender auf ihr genaues Alter hin zu untersuchen. Fallweise treten daher Attribuierungen auf, die einen Codex bloß ganz allgemein als „alt“ oder als „sehr alt“ bezeichnen. Auf genauere Altersangaben wären folglich Eckharts Quelleneditionen zu untersuchen. Als entsprechende Eigenschaftsworte in den Briefen treten *vetus* (zwei Mal in der Grundstufe: BE 22; BE 27) oder *vetustus* auf, letzteres Adjektiv zwei Mal in der Grundstufe (BE 15; BE 21) und zwei Mal als Superlativ (BE 8; BE 23). Fünf Belegstellen lassen sich für Formen von *antiquus* ausfindig machen, nicht aber in Verbindung mit *codex*.

Angaben zum Material, zur physischen Beschaffenheit bzw. zum Beschreibstoff erwähnter Textträger kommen an den untersuchten Stellen fast nicht vor. Nur fünf Mal werden Pergamentcodices (oder Fragmente daraus) als solche erwähnt, zwei Mal explizit in Verbindung mit dem Adjektiv *membranaceus* (BE 14; BE 17) – die Verwendung dieses Adjektivs beschränkt sich auch auf diese beiden Fälle –, einmal als Buch (*liber*) *in membrana* (BE 31), einmal implizit als *unus in membrana* (BE 33) und einmal indirekt als Vorlage für eine Abschrift (*Mitto tamen Sermonem beati Volcuini ex veteri membrana descriptum* – BE 32). Dass die Wendung *in membrana* dabei einen ledernen Einband meint, ist im Falle der untersuchten Stellen auszuschließen.

Der Begriff *charta* (nur in dieser Schreibweise, nicht als *carta*) ist in den Eckhart-Briefen insgesamt neun Mal anzutreffen (BE 59–67), allerdings in ganz unterschiedlichen Bedeutungen (z.B. für geographische Karten, für Spielkarten

oder für Papier, das als Rohmaterial unterschiedlicher Qualität für gedruckte Werke oder für Handschriften dient), nie explizit in Verbindung mit dem Terminus *codex* und nur einmal mit schwachem, da sehr allgemeinem Bezug auf Textträger in Klöstern (und scherzhaft-ironisch): *Per integrum fere annum nunc huc nunc illuc vagatus excussi monasteriorum nostrorum chartas pulvere obsitas* (BE 67 = 34). Angaben zum Format von Überlieferungsträgern treten in Form der Wendung *in folio* vier Mal auf, zwei Mal in Verbindung mit *tomus* (BE 35; BE 37), zwei Mal zusammen mit *volumen* (BE 17; BE 36). Die Wendung *in octavo* ist einmal vertreten (BE 31), ein zweites Mal wird sie umschrieben mit *forma, quam octavam vocant* (BE 1).

Folgende Zeitworte erscheinen in Verbindung mit einer Form von *codex* (hier im Infinitiv wiedergegeben in alphabetischer Reihenfolge; insgesamt sind es 29): *admirari, avehi, carere (non posse), comparere, continere, custodire, desiderare* (2 Mal), *detegere, emere, esse* (2 Mal), *habere, habere velle, invenire, mereri* (scil. *edi et conferri*), *mittere, perlustrare, pertinere, placere, proferre* (scil. *editionem codicis*, 2 Mal), *recipere* (2 Mal), *reperire, venire, videre* (2 Mal).

Mit *videre, admirari* und *placere* sind dabei drei Verben verhältnismäßig häufig vertreten, die mit einer Sinneswahrnehmung Eckharts in Verbindung stehen oder unmittelbar auf diese verweisen – Codices wurden während seiner Reisen bewusst in Augenschein genommen. *Detegere, reperire* und *invenire* stehen für die Auffindung eines Überlieferungsträgers. Auch weisen verhältnismäßig viele Verben dieser Reihe auf den regen postalischen Verkehr hin, dessen sich Gelehrte bedienten, um untereinander Codices, Auszüge daraus oder Abschriften auszutauschen (*mittere, recipere, venire*). Die Edition einer Quelle bzw. die editorische Tätigkeit wird zwei Mal mit *proferre* umschrieben. Auch *comparere* steht in diesem Zusammenhang. Mit *carere non posse* und *custodire* wird darauf hingewiesen, dass man in einzelnen Aufbewahrungsstätten großen Bedacht auf die Bestände nahm. Eckhart wurde eine Einsichtnahme, Exzerpierung oder gar eine Entlehnung von Überlieferungsträgern fallweise erschwert bzw. verunmöglicht. Nicht immer gelangte der Gelehrte an jene Objekte, die er gerne untersucht hätte (*habere velle*).

Im Corpus der untersuchten Pez-Briefe, welche zwischen 1709 und 1715 verfasst wurden, findet man den Begriff *codex*, wie bereits erwähnt, an 23 Stellen (BP 1–

BP 23). Gemessen an der Gesamtzahl der in diesem Corpus enthaltenen sprachlichen Einheiten (23.414 Tokens insgesamt) verwenden die Gebrüder Pez den Begriff *codex* somit weit seltener in ihrer aktiven Korrespondenz als Eckhart. Hinsichtlich seines *Thesaurus anecdotorum novissimus*²²⁴, im Druck erschienen zwischen 1721 und 1729, konnte von der Forschung bereits gezeigt werden, dass Bernhard „den Begriff *codex* in zweifacher Weise (und ohne darauf hinzuweisen) verwendet. Während er damit einerseits eine physische Einheit in Form einer (meist mittelalterlichen) Handschrift benennt, die in den meisten Fällen auch die Basis der von ihm durchgeführten Edition darstellt, bezeichnet er [damit] andererseits [...] in der Edition von ihm selbst zusammengestellte Quellensammlungen bzw. unter dem Titel *codex diplomaticus* eine Sammlung von Urkunden. [...] Als bestes Beispiel einer von Bernhard Pez zusammengestellten Quellensammlung in Verbindung mit dem Gebrauch des Begriffes *codex* kann der sechste Band des *Thesaurus* herangezogen werden.“²²⁵ Dessen Titel lautet: *Codex historico-diplomatico-epistolaris*.

Für das untersuchte Briefcorpus aus den Jahren 1709 bis 1715 trifft dieser Befund nicht zu. Der Begriff *codex* wird dort *ausschließlich* für handschriftliche Überlieferungsträger verwendet, in mindestens elf von dreiundzwanzig Fällen sogar explizit in Verbindung mit dem Adjektiv *manuscriptus* (BP 1; 3–8; 10; 17; 19; 21). Aus dem Kontext der übrigen Belegstellen, bei denen dieses Adjektiv nicht auftritt, lässt sich erschließen, dass ebenfalls handschriftliche Überlieferungsträger gemeint sind, nicht etwa frühneuzeitliche Drucke oder selbst angelegte bzw. selbst edierte Quellensammlungen.

Die Korrespondenz zwischen Johann Georg Eckhart und den Brüdern Pez setzt zur Jahreswende 1717/1718 ein. Für ein vollständiges Bild darüber, wie die beiden Melker Benediktiner den Begriff *codex* im Laufe ihrer Forschungstätigkeit gebrauchten, müssten weitere Untersuchungen angestellt werden, die an dieser Stelle nicht geleistet werden können. Auf Grund der hier analysierten Quellenbasis lässt sich jedoch vermuten, dass gerade in der Verbindung *codex manuscriptus* nicht nur die bereits beobachtete Verwendung für *handschriftliche* Überlieferungsträger angelegt ist, sondern gerade auch jene für gedruckte

²²⁴ Vgl. hierzu e.g. Wallnig, Briefe v. J.G. Eckhart an B.Pez (wie Anm. 15), 64.

²²⁵ Manuela Mayer, Das Chartular von St. Emmeran und seine Edition durch Bernhard Pez. In: MIOG 125/2 (2017), 288–303; hier: 301 und 302.

Quellensammlungen und für Editionen. Das Adjektiv *manuscriptus* dient hier offenbar zur Distinktion. Es müsste demnach das gesamte Pez'sche Oeuvre dahingehend untersucht werden, ob Bernhard mit dem Begriff *codex* vornehmlich dann Editionen meint, wenn er das Adjektiv *manuscriptus* weglässt, was in den hier untersuchten Briefen jedoch, wie gesagt, nicht zutrifft.

Für das sieben Mal belegte Nomen *manuscriptum* (BP 24–30) gilt derselbe Befund wie im Falle der Eckhart-Briefe: Es lässt sich semantisch nicht genau feststellen, wo die Trennlinie zwischen ihm und *codex* verläuft. Es steht jedoch zu vermuten, dass damit auch lose Blätter mit Handschriften gemeint sein können sowie Fragmente. Seltener als Eckhart verwendet Pez den Begriff *codex* mit explizitem Bezug auf einen konkreten Überlieferungsträger. Das zeigt sich insbesondere darin, dass der Singular bei Eckhart, wie bereits gesagt, 23 Mal belegt ist, der Plural jedoch nur sieben Mal. Nachgerade umgekehrt verhält es sich mit den Belegstellen im Pez-Corpus: Acht Vorkommen im Singular (BP 2; 9; 11; 12; 14; 16; 18; 23) stehen 15 Pluralformen gegenüber. Öfter als für konkrete Überlieferungsträger gebraucht der Benediktiner den Begriff als Chiffre in einem Kontext, der grosso modo auf seine eigene gelehrte Tätigkeit im Allgemeinen hinweist, auf seinen eigenen Umgang mit Codices in Bibliotheken und auf jenen in der *Res publica litteraria* generell, kurz: auf die historisch-kritische Tätigkeit seiner selbst und auf jene seiner Kollegen.

Angaben zum Inhalt der Überlieferungsträger finden sich im unmittelbaren Kontext von zumindest sechs Erwähnungen (BP 1; 9; 14; 16; 21; 23), solche zum Aufbewahrungsort explizit bei ebenfalls sechs Belegstellen (BP 1; 2; 7; 8/9; 16; 18). Ferner ist darauf hinzuweisen, dass Pez sich zweimal höchst negativ in Bezug auf Aufbewahrungsstätten äußert, und zwar insofern, als er im Chorherrenstift Herzogenburg und in St. Andrä an der Traisen offenbar kaum handschriftliche Bestände ausfindig machen konnte, die für seine Forschungen von Interesse gewesen wären (BP 4 & 5).

Vier Mal findet man Angaben zum Alter der angesprochenen Codices. In der Regel trachtet Pez hier nach genaueren Angaben als Eckhart: *anno 1606* (BP 2), *saeculo XV*. (BP 5), *saeculi noni* (BP 9), *coaevo* (scil. *saec. XIII*; BP 18). Ob diese Zuschreibungen auch korrekt sind, kann im Rahmen der vorliegenden Arbeit leider nicht untersucht werden. Auf allgemeine Attribuierungen ohne genaue

Angaben hinsichtlich Alter weicht Pez nur zwei Mal aus: *vetustus* und *vetus* sind in Verbindung mit dem Begriff *codex* jeweils nur einmal im Corpus belegt (BP 1; BP 20). Ausgesprochen auffällig ist, dass der um Contenance und um gewählten Ausdruck sonst so bemühte Benediktiner sich an einer der untersuchten Stellen höchst abfällig über Handschriftenbestände äußert, da diese offenbar nicht seinen Erwartungen entsprachen: *Codices manuscripti* aus dem Chorherrenstift Herzogenburg bedenkt er mit den Adjektiven *communes*, *passim vulgati* und *denique proletarii* (BP 4).

Angaben zum Material und zur physischen Beschaffenheit von *Codices* treten bei Pez noch spärlicher in Erscheinung als bei Eckhart, zumindest dann, wenn man nach den Stichworten *charta* oder *membrana* im Corpus sucht bzw. nach den davon abgeleiteten Adjektiven *membranaceus* und *chartaceus*. Für den Wortstamm **membran-* findet sich keine einzige Belegstelle. Das Nomen *charta* ist insgesamt fünf Mal anzutreffen (BP 31–35), wobei das Bedeutungsspektrum wiederum variiert. Das Adjektiv *chartaceus* fehlt gänzlich. Den Terminus *charta* scheint Pez im untersuchten Briefcorpus – anders als erwartet – nie für den Beschreibstoff Papier zu verwenden. Vielmehr bezeichnet er damit entweder Urkunden (in vier von fünf Fällen: BP 31–33 & 35) oder Landkarten (eine Belegstelle: BP 34). Ein Konnex zum Terminus *codex* lässt sich nicht nachweisen. Angaben zum Format von Überlieferungsträgern (einschließlich gedruckten Editionen) finden sich fünf Mal in Form der Wendung *(in) folio* (BP 36–40) und ebenso fünf Mal in Form der Wendung *(in) quarto* (BP 41–45).

Das Oktav-Format wird vier Mal genannt (*in octavo*; BP 46–49) – und zwar ausschließlich in Zusammenhang mit Editionen.

Folgende Verben verwendet Pez in Verbindung mit dem Begriff *codex* (Wiedergabe wiederum im Infinitiv in alphabetischer Reihenfolge; Gesamtzahl: 30): *accipere* (scil. *nullos!*), *asportare*, *commendare*, *complecti*, *efferre* (zwei Mal), *esse* (drei Belege; zwei Mal elliptisch ausgelassen), *evolvere*, *excutere* (vier Mal), *exscribere* (zwei Mal), *extare*, *facessere*, *habere* (zwei Mal), *incidere*, *legere*, *mittere* (zwei Mal), *notare*, *recipere*, *reperire*, *scribere* (2 Mal), *videre*.

Obgleich der Begriff *codex* im untersuchten Pez-Corpus also seltener auftritt als in den Eckhart-Briefen, ist die Anzahl der damit verbundenen Verben geringfügig höher als dort, was in der Struktur der Sprache Bernhards begründet scheint.

Semantisch betrachtet beziehen folgende Verben sich auf den physischen Umgang mit den Überlieferungsträgern in ihren Aufbewahrungsstätten sowie auf den Vorgang der Entlehnung und auf den postalischen Verkehr: *accipere*, *asportare*, *efferre*, *mittere*, *recipere*. Auch *facessere* ist in diesen Zusammenhang einzuordnen: Die Aussage *Codex ... eruditorum votis negotium facessit* (BP 14) spiegelt den Umstand, dass ein aus der Hofbibliothek entliehener Codex nur noch mit Mühe an die ursprüngliche Stätte seiner Aufbewahrung zurückzuholen war.

Auf das physische Vorhandensein eines Überlieferungsträgers weisen *esse* und *extare*. In Zusammenhang mit gelehrter Kommunikation stehen *commendare* und *notare*. Das Ereignis einer (zufälligen) Auffindung spiegelt sich in den Verben *incidere* und *reperire*. Auf den Inhalt von Überlieferungsträgern bezieht sich *complexi*.

Gelehrte Tätigkeiten in Zusammenhang mit Codices werden – durchaus scherzhaft-ironisch – im Allgemeinen mit *evolvere* und mit *excutare* (scil. *pulvere*) umschrieben. Spezifischere und genauer bezeichnete Tätigkeiten gelehrten Charakters sind dagegen *exscribere*, *scribere* und *legere*.

Der Begriff *liber* (Buch) ist in den Eckhart-Briefen 14 und im Pez-Corpus 27 Mal belegt (BE 38–51; BP 50–76). Da die Quantitäten von Vokalen bei der rein linguistischen Annotation nicht berücksichtigt werden konnten, bedarf es während der Corpus-Abfrage erhöhter Aufmerksamkeit sowie der Berücksichtigung des Kontextes und der PoS-Tags. Ohne Kennzeichnung der Vokalquantitäten stellt *liber* nämlich eine Homographie dar, bei klassisch-lateinischer Aussprache jedoch keine Homophonie. Die Quantität des *i* ist bedeutungsunterscheidend. Dies gilt es bei der Corpus-Abfrage besonders zu beachten. Mit langem *ī* handelt es sich bei klassischer Aussprache nämlich um jenes Adjektiv, das in deutscher Übersetzung *frei* bedeutet, mit kurzem *i* bezeichnet die Zeichenkette ein Buch, was jedoch nicht gleichzusetzen ist mit *codex* (zumindest nach antikem Verständnis). Eine Möglichkeit, Vokalquantitäten im TokenEditor zu annotieren, gibt es derzeit noch nicht. Unabhängig davon war die maschinelle Zuordnung der Wortarten-Tags (ADJ gegenüber NN) meist korrekt.

Ein strikt einheitlicher Gebrauch des Begriffes *liber*, etwa in strenger Abgrenzung zu *codex*, lässt sich in den Eckhart-Briefen nicht feststellen. Eher liegt ein gemischter Gebrauch vor. Dabei zeigen sich folgende Tendenzen: (1) In der

überwiegenden Anzahl aller Fälle verwendet Eckhart den Terminus für gedruckte Textträger (sechs eindeutige Belegstellen: BE 40, 42, 44, 45, 46, 51). (2) In nur einem einzigen lässt sich mit Sicherheit sagen, dass ausschließlich *ein* konkreter *handschriftlicher* Codex gemeint ist (BE 41). An dieser Stelle ist *liber* demnach weitestgehend synonym mit *codex*, wie Eckhart letzteren Begriff überwiegend verwendet. (3) Im Falle von sechs Belegstellen (BE 38, 39, 43, 47, 48, 50) ist nicht eindeutig abzugrenzen, ob gedruckte oder handschriftliche Textträger gemeint sind. Den Terminus *liber* verwendet Eckhart hier vielmehr als Sammel- bzw. als Überbegriff, der beides umfassen kann. Bei diesen Belegen ist es zumindest höchst wahrscheinlich bzw. keinesfalls auszuschließen, dass *auch* handschriftliche Codices mit gemeint sind. An einer dieser Stellen steht *codex* sogar in unmittelbarer sprachlicher Nachbarschaft im selben Satz (BE 38). Diese Stelle bestätigt explizit, dass *liber* in den Eckhart-Briefen als Über- bzw. als Sammelbegriff anzusehen ist, dem der Terminus *codex* untergeordnet sein kann. Letzterer wird, wie weiter oben bereits gezeigt, zumeist für handschriftliche Überlieferungsträger gebraucht.

Der Diminutiv *libellus* findet sich in den Eckhart-Briefen sieben Mal (BE 52–58). Auffällig und nicht zuletzt überraschend in diesem Zusammenhang ist, dass er semantisch von *liber* abzuweichen scheint – und das nicht allein auf Grund der Tatsache, dass es sich um einen Diminutiv handelt. Mit *codex* ist er nicht gleichzusetzen. Eckhart verwendet *libellus* insgesamt sechs Mal in Verbindung mit inhaltlichen Angaben hinsichtlich der bezeichneten Werke (BE 52–56 & BE 58). Fallweise scheinen diese inhaltlichen Angaben (kanonischen?) Werktiteln (?) lateinischer Sprache zu ähneln, wie sie in der Antike und während des Mittelalters sehr verbreitet waren und durch die auf den Inhalt eines Textes geschlossen werden kann. Gemeint sind Werktitel bestehend aus der Präposition *de*, gefolgt von einem Substantiv im Ablativ (BE 52: *de necromantia libellus*; BE 53: *ineditum de amore libellum*; BE 56: *libellum de numis Wirceburgensibus* et c.).

Zwar steht der Terminus *liber* an manchen Belegstellen ebenso in Zusammenhang mit inhaltlichen Angaben bezüglich der damit bezeichneten Bücher, doch ist dieser Zusammenhang meist lockerer als bei *libellus*. Die Wendung *liber de ...* findet sich an keiner einzigen Stelle. Der Begriff *codex* meint im Gegensatz dazu den Überlieferungsträger als physisch-materiellen Gegenstand, der Werke, Texte,

also *libros* bzw. *libellos* – die „eentlichen“ Träger des Inhaltes – enthalten kann. Obgleich, wie bereits gezeigt, inhaltliche Angaben (oft recht allgemeiner Natur) in Verbindung mit dem Terminus *codex* nicht selten sind, existiert die Wendung *codex de ...* in den Eckhart-Briefen nicht.

Eher wird *codex* in Verbindung mit Genetiv Plural verwendet (z.B. BE 7: *codex epistularum*; BE 16: *codex capitularium legumque veterum*), in Verbindung mit einem Relativsatz (BE 5: *codex, qui continet...*), mit einem Präsenspartizip (BE 23: *codex...continens*) oder in Verbindung mit einem Adjektiv (z.B. BE 9: *Theotiscum!*), um den Inhalt summarisch zu beschreiben, fallweise besonders in Hinblick auf die enthaltenen Text-Genera. Der Terminus bzw. die Wendung *libellus (de)* bezieht sich hingegen *nie* auf Text-Genera, sondern oftmals auf den Werkinhalt.

Im untersuchten Pez-Corpus wird der Begriff *liber* an insgesamt fünf Belegstellen für Einzelbände der geplanten *Bibliotheca Benedictina* verwendet (BP 50; 52–55). In *tredecim fere libris* würden alle benediktinischen Schriftsteller umfasst, so Bernhard in mehreren seiner frühen Rundschreiben. Einmal bezieht *liber* sich auf Bernhards gedruckten Brieftraktat *Epistolae apologeticae pro ordine S. Benedicti* (BP 76).

Verallgemeinernd gesprochen erscheint die Semantik des Begriffes *liber* bei Bernhard insgesamt unscharf. An fünfzehn Stellen ist er weitestgehend synonym mit unserem Terminus *Buch* bzw. *Werk*, einschließlich jener Stelle, an der von den *Epistolae apologeticae* die Rede ist (BP 51; 57; 58; 61–63; 65; 67–70; 72–74; 76). Häufig wird er auch verwendet, um eine Unterkategorie eines größeren Werkes zu bezeichnen, das aus mehreren Teilen besteht. *Liber* kann somit synonym sein mit unserem Terminus *Band* (engl. *volume*), wofür sich im untersuchten Pez-Corpus elf Belegstellen finden (BP 50; 52–56; 59; 60 [?]; 66; 71; 75). Darin enthalten sind auch jene erwähnten Passagen, in denen der Melker Historiker von seiner *Bibliotheca Benedictina* schreibt. Dieser synonyme Gebrauch – *liber* = *Buch* = *Band eines mehrteiligen Werkes* – ist bereits in der klassischen Antike üblich. Auch in den modernen Altertumswissenschaften ist er heute nach wie vor omnipräsent. Von der physischen Erscheinungsform eines Codex ist dabei jedoch weitestgehend abzusehen. Bis in die Spätantike war *liber* überwiegend synonym mit *volumen*, der (Buch-)Rolle aus Papyrus.

Wie schon im Falle der Eckhart-Briefe kann also auch im untersuchten Pez-Corpus von einem gemischten Gebrauch des Terminus *liber* gesprochen werden. Für Bernhard scheint es dabei nur eine untergeordnete Rolle zu spielen, ob die damit bezeichneten Überlieferungsträger gedruckt sind, noch gedruckt werden oder nur handschriftlich vorliegen. Eine scharfe Trennlinie kann hier nicht gezogen werden. Bernhard scheint den Begriff *liber* vornehmlich dann zu gebrauchen, wenn es ihm *nicht* darauf ankommt, auszusagen, ob ein Überlieferungsträger handschriftlich oder gedruckt vorliegt oder noch vorliegen wird. Ganz anders verhält es sich, wie wir gesehen haben, mit dem Terminus *codex*, den Bernhard in den untersuchten Briefen ausschließlich für handschriftliche Textträger verwendet, was überwiegend, nicht aber ausschließlich, auch in den Eckhart-Briefen zu beobachten ist. *Codex* ist mit *liber* also nicht synonym. Dass *liber* bei Pez *codex* übergeordnet wäre, lässt sich ebenso wenig bestätigen. Ein mehrbändiges Werk (aus mehreren *libri*) kann zwar in Form mehrerer Codices überliefert sein, doch diese Verbindung findet sich im untersuchten Pez-Corpus nicht. Jene Stellen, an denen der Melker Geschichtsforscher *liber* im Sinne von *Band* oder *Buch* verwendet und sich dabei auf Werke bezieht, die nicht von ihm stammen, lassen zwar den Schluss zu, dass er handschriftliche Überlieferungsträger meinen könnte. Einigermassen verifizieren lässt sich das jedoch nur an zwei Stellen: Dort tritt das Adjektiv *manuscriptus* in direkter Verbindung mit *liber* auf (BP 2 = 64; 66). Dass mit *liber* dabei der *einzelne* physische Codex gemeint ist, ist an diesen beiden Stellen zwar höchst wahrscheinlich. Mit letzter Gewissheit steht dies aber nicht fest, denn ein *liber* kann sich, obgleich *manuscriptus*, kraft Überlieferung auch auf mehrere Codices verteilen.

Den Terminus *libellus*, der im untersuchten Pez-Corpus insgesamt sieben Mal belegt ist (BP 77–79; 81–84), verwendet der Benediktinerhistoriker überwiegend für gedruckte Werke (sechs Mal: BP 1=78; 79; 81–84), einmal wahrscheinlich Mal für einen handschriftlichen Überlieferungsträger (BP 77). Semantisch ist der Begriff damit weitestgehend identisch mit *liber*, also mit *Buch* oder mit *Werk* im Sinne eines Buches. Die Untereinheit eines größeren Werkes im Sinne von *Band* wird zwei Mal mit *libellus* bezeichnet (BP 77; 81).

Den doppelten Diminutiv *libellulus* verwendet Bernhard Pez nur einmal (BP 80), und zwar im Nominativ Plural. Er bezeichnet damit eines seiner eigenen

Frühwerke (*De irruptione Bavarica in Tirolim*), bestehend aus drei Bänden. Der doppelte Diminutiv ist dabei der Bescheidenheitstopik geschuldet.

Weit seltener als der Terminus *liber* bzw. als dessen Verkleinerungsformen tritt der Begriff *volumen* in den untersuchten Textcorpora auf. Bei Eckhart ist er insgesamt sechs Mal belegt (BE 68–73), bei Pez fünf Mal (BP 85–89). Beide Gelehrten verwenden ihn weitestgehend in der Bedeutung unseres Wortes *Band*. Der Begriff scheint dabei weder auf handschriftliche noch auf gedruckte Überlieferungsträger beschränkt zu sein. Vielmehr kann er für beides verwendet werden. Pez gebraucht ihn häufiger für edierte Werke (drei Mal: BP 85; 87; 89), nur einmal für einen handschriftlichen Textträger (BP 86 – hier ist *volumen* offenbar synonym mit *codex*). Auch Eckhart verwendet *volumen* sowohl für gedruckte Werke (BE 69 & 73) wie für handschriftliche (BE 68; BE 71=17; BE 72=36).

An insgesamt zwei Stellen (bei beiden Historikern zusammen) ist nicht klar, ob damit handschriftliche oder gedruckte Texte gemeint sind (BE 70=47; BP 88). Ganz allgemein gesprochen fällt auf, dass damit häufig Zahlangaben einhergehen, aus wie vielen Bänden ein Gesamtwerk besteht. In Verbindung mit Adjektiven wie *spissus* (BE 68), *magnus* (BP 85), *grandis* (BP 86) oder *vastus* (BP 87) wird dabei häufig auf die beträchtliche physische Ausdehnung der Überlieferungsträger verwiesen.

2.6) Schlussbetrachtung

Die vorangegangenen Untersuchungen haben vor allem eines deutlich gezeigt, nämlich den Wert der Lemmatisierung innerhalb annotierter Corpora. Zumal für Corpus-Abfragen, wie sie hier durchgeführt wurden, stellt sie *die* Grundvoraussetzung schlechthin dar. Einer ihrer Vorteile besteht darin, dass sie unter zahlreichen Teilschritten linguistischer Annotation zu jenen Prozessen zählt, die relativ einfach durchzuführen sind. Zwar ist auch bei der Lemmatisierung stets auf den sprachlichen Kontext zu achten, wie im Übrigen bei jedem anderen Teilschritt linguistischer Annotation auch, doch stellt sie an Annotatorinnen und Annotatoren geringere Anforderungen verglichen mit der morphologischen Annotation. Kaum ein linguistisch annotiertes Corpus wird daher ohne Lemmatisierung auskommen. Als äußerst hilfreich hat sich auch die Kennzeichnung von Wortarten (das PoS-

Tagging) erwiesen. Ganz abgesehen davon, dass ein Corpus mit jeder Unterart linguistischer Annotation für sprachwissenschaftlich feinere digitale Abfragen geöffnet wird, zeigte sich der Wert des PoS-Taggings beispielsweise in Hinblick auf Homographien (*liber!*). Erhellend war in diesem Zusammenhang, dass zeitgemäße Tagging-Tools mit diesem Phänomen bereits problemlos zurechtkommen.

Von der rein quantitativen Ausgabe der Tagging-Ergebnisse kann auf diesem Wege leicht auf deren Kontext fokussiert werden, was wiederum die Grundvoraussetzung für semantische Analysen darstellt, wie sie zuvor durchgeführt wurden. An vorangegangene Forschungsergebnisse²²⁶ konnte hierbei ergänzend und differenzierend angeknüpft werden.

Nicht ohne ein gewisses Moment der Überraschung zeigte sich hierbei, dass etwa die Verwendung und die Semantik ein und derselben Zeichenkette bei ein und demselben Autor durchaus variieren können, je nach dem, in welchem Kontext er sie verwendet, z.B. *codex* bei Pez in Abhängigkeit davon, in welchem Medium er den Terminus gebraucht (Brief gegenüber Edition).

Nicht zuletzt zeigte sich erwartungsgemäß, dass der Gebrauch bestimmter Termini auch von Autor zu Autor variieren kann (siehe etwa *libellus*). Der Verfasser vorliegender Zeilen sieht damit einmal mehr zwei Phänomene bestätigt, die innerhalb der Digital Humanities bereits wiederholt beobachtet wurden, nämlich erstens, dass digital unterstützte Forschung einerseits dabei hilft, durchaus überraschende, unvermutete Ergebnisse zu Tage zu fördern, dass aber, zweitens, auch bestehende Vermutungen vielfach durch digitales Arbeiten bestätigt werden. Die Herausforderung besteht seiner Ansicht nach darin, die zahlreichen digitalen Editionen und Textsammlungen, die in jüngster Zeit mit oder auch ohne linguistische Annotation zur Verfügung gestellt wurden, tatsächlich als Quellenmaterial auch auszuwerten. Die Forschung würde vermutlich um ein Vielfaches bereichert, wenn digitale Corpora mit demselben Aufwand beforscht würden, mit dem sie erstellt werden. Das Austrian Baroque Corpus stellt in diesem Zusammenhang ein höchst positives Beispiel für ein gut beforschtes Corpus dar. Ihm kommt insofern Vorbildfunktion zu.

²²⁶ Vgl. Mayer, Chartular (wie Anm. 225).

Im denkbar kleinsten Rahmen ist dies im Zuge der vorliegenden Studie geschehen. Sie soll hiermit beschlossen sein. Denn längst übersteigen Aufwand und Auswertungspotential, wie sie etwa mit der Anlage flächendeckender morphologischer Annotation einhergehen, die gegebenen Möglichkeiten.

In den allermeisten Fällen wird eine digitale Edition mit einer linguistischen Basisannotation auskommen, also mit PoS-Tagging und mit Lemmatisierung. Morphologische Annotation ist vor allem unter dem Aspekt verfeinerter linguistischer Such- und Forschungsanfragen zu erwägen.

Der Aufwand, der mit flächendeckender morphologischer Annotation einhergeht, sollte jedoch nicht unterschätzt werden. Insofern ist auf morphologischer Ebene an Stelle einer flächendeckenden, vollständigen Annotation an eine spezifische Kennzeichnung zu denken, die ganz gezielten, vorab abgeklärten Suchanfragen entgegenkommt. Erste Auswertungen kann man auch schon während des Annotationsprozesses selbst vornehmen, da dieser mit einer sehr intensiven Beschäftigung mit der Textgrundlage einhergeht. Insofern fördert Annotation die feingranulare Beschäftigung mit den Texten unserer Quellen.

Als Hilfestellung für künftige digitale Corpora bzw. Editionen sollen auf Basis all des bisher Gesagten am Schluss dieser Arbeit drei Empfehlungen ausgesprochen werden – als Zusammenfassung im Sinne des NLP lateinischer Texte für die Editionswissenschaft. Diese Empfehlungen sollen bei der Beantwortung der Fragen helfen, (1) welche Arten linguistischer Annotation, (2) welches Tagset und (3) welche Tagger bei der Annotation lateinischer Textcorpora zukünftig verwendet werden könnten.

(1) Linguistische Basisannotation mit Wortarten-Tagging und Lemmatisierung stellt einen äußerst gangbaren, weil weniger aufwendigen Weg als Treebanking dar, Textcorpora und digitale Editionen für die sprachwissenschaftliche Forschung zu öffnen.

Im Gesamtkontext linguistischer Annotationsarten sind diese beiden Vorgänge geradezu unverzichtbar. Gleichzeitig lassen sie, da relativ einfach durchzuführen, ausreichend Raum für zusätzliche Annotationen, auch nicht streng linguistischer Art, z.B. für das Tagging von Named Entities, also von Orten, Personen, Organisationen und Werken.

Die aufwendigere morphologische Annotation stellt in Hinblick auf sprachgeschichtliche Fragestellungen eine durchaus reizvolle Option dar. Ob sie flächendeckend oder in spezieller Art und Weise an einem Corpus durchzuführen ist, wird von Fall zu Fall zu entscheiden sein, abhängig von den konkreten Forschungs- und Abfragebedürfnissen innerhalb konkreter Projekte.

Flächendeckendes Treebanking ist nur dann durchzuführen, wenn ausdrücklich syntaktische Fragestellungen im Vordergrund der Forschung stehen. Diesen Fragestellungen kann man auch bei der PoS-Annotation einen Schritt entgegenkommen, wenn man den Versuch unternimmt, morphosyntaktische Phänomene durch Tagsets zu explizieren, die ihrem Grundgedanken nach für die Kennzeichnung von Wortarten konzipiert wurden.

(2) Der Verfasser vorliegender Arbeit muss sich leider eingestehen, gegenwärtig mit keinem einzigen der verfügbaren Tagsets für die lateinische Sprache vollständig zufrieden zu sein. Das im Anschluss zu dieser Arbeit konzipierte Tagset wird den Versuch darstellen, hier vorübergehend Abhilfe zu schaffen (erscheint 2020 im Rahmen des Sammelbandes *Digitale Edition in Österreich*, herausgegeben vom Kompetenznetzwerk Digitale Edition). Gleichwohl versteht es sich keineswegs als „Goldstandard“. Es ist daher nicht auszuschließen, dass es in Zukunft weiteren Bearbeitungen und Korrekturen unterzogen werden muss.

(3) Aus der hier skizzierten Situation heraus, dass sich im Bereich des lateinischen NLP während der letzten Jahre und Jahrzehnte eine höchst heterogene, dezentrale Landschaft unterschiedlicher, teils ausreichender, teils ungenügender Tagsets entwickelt hat, ergibt sich ein Problem. Es gibt, zumal für das Tagging neu-lateinischer Texte, gegenwärtig keinen einheitlichen Standard annotierter Daten. Weder in Hinblick auf Tagsets noch in Hinblick auf ein automatisches Tagging-Tool kann derzeit eine einheitliche, letztgültige Empfehlung ausgesprochen werden. Vielmehr müssen zunächst die folgenden Frage gestellt und beantwortet werden: (I) Auf welches Tagset ist ein Tagging-Tool gegenwärtig trainiert und ausgelegt? (II) Genügt das Tagset den Ansprüchen der Forschung? (III) Welche automatischen Trefferquoten liefert ein in Frage kommender Tagger X mit dem Tagset Y?

Wenn vor allem die Fragen (II) und (III) in zufriedenstellender Weise beantwortet sind, so kann guten Gewissens auf ein bestehendes Setting (Tagset und Tagger) zugegriffen werden, um eine automatische linguistische Annotation durchzuführen.

Ist dies nicht der Fall, das heißt: sind moderne Tagger grundsätzlich nicht auf Tagsets ausreichender Qualität für die lateinische Sprache ausgelegt, so gibt es immer noch die Möglichkeit, Tagger neu auf ein eigens erstelltes Tagset zu trainieren. Das setzt allerdings die Erstellung eines händisch annotierten Trainingscorpus voraus – und der Aufwand dieses Vorgangs ist nicht zu unterschätzen. In so einem Fall stellt sich zudem die Frage, für welchen Tagger man sich entscheidet – und zwar *ohne* im Vorhinein mit letzter Sicherheit über die Qualität der automatisch zu generierenden Annotationsergebnisse Bescheid zu wissen, welche aus dem Zusammenwirken von Tagger und Tagset zustande kommen. Hier scheint es sich zu empfehlen, auf die bereits genannten Tagger zurückzugreifen: FLORS, Lapos oder MarMot. Diese konnten unter bisherigen Tests mit der deutschen sowie mit der lateinischen Sprache bereits äußerst ansehnliche Ergebnisse erzielen, wenngleich mit etwas reduzierten lateinischen Tagsets. Ihre Performance stimmt zuversichtlich, dass die Qualität der Annotationsergebnisse auch nach erneutem Training mit leicht verbesserten Tagsets ein äußerst akzeptables Niveau aufweisen wird.

Freilich ist dabei stets auf neuere technologische Entwicklungen, auf neuere und auf möglicherweise bessere Tools aus dem Bereich des NLP zu achten. Besonderes Augenmerk gilt es auf ein erst jüngst (2018) von Marco Passarotti ins Leben gerufenes Projekt an der Università Cattolica del Sacro Cuore in Mailand, einer Wirkungsstätte Roberto Busas, zu legen: „LiLA“ – „Linking Latin“. Von diesem Projekt sind bis spätestens 2023 wesentliche Verbesserungen im Bereich des lateinischen NLP zu erwarten. Hierauf sei an dieser Stelle abschließend verwiesen:

„Despite the headway made in the last decade in building, sharing and exploiting linguistic resources and tools for the automatic processing of Latin, these remain incompatible. The objective of LiLa (2018–2023) is to connect and ultimately exploit the wealth of linguistic resources and NLP tools for Latin created so far, in order to bridge the gap between raw language data, NLP and knowledge

descriptions. To do so, LiLa is building an open-ended Knowledge Base using the Linked Data paradigm, concurrently adding Latin to the multilingual Linguistic Linked Open Data ([LLOD](#)) cloud.²²⁷

„LiLa (a) enhances a large amount of Latin texts with PoS-tagging and lemmatisation, (b) harmonises the annotation of the three Universal Dependencies treebanks for Latin, (c) improves the lexical coverage of the *Latin WordNet* and the valency lexicon *Latin-Vallex*, and (d) expands the textual coverage of the *Index Thomisticus Treebank*. Furthermore, LiLa builds a set of newly-trained models for PoS-tagging and lemmatisation, and works on developing and testing the best performing NLP pipeline for such a task.“²²⁸

Die Anlage neuer Treebanks innerhalb von LiLA ist nicht geplant, wohl aber die Vereinheitlichung bestehender.²²⁹ Es bleibt abzuwarten, ob LiLA dazu beitragen kann, die Möglichkeiten der automatischen syntaktischen Annotation wesentlich zu verbessern. Nach Auskunft von Giuseppe G.A. Celano (Universität Leipzig) liegen die Trefferquoten elektronischer Tools, die Treebanks erstellen können, bei nur rund 60% (Stand Frühjahr 2018).²³⁰

²²⁷ <https://lila-erc.eu/#about-1>; letzter Zugriff: 28.05.2019.

²²⁸ <https://lila-erc.eu/about/>; eine vorläufige Publikationsliste zu LiLA findet sich unter: <https://lila-erc.eu/output/>; letzter Zugriff auf beide URLs: 28.05.2019.

²²⁹ Marco Passarotti, dem Initiator und Leiter von LiLA, sei an dieser Stelle herzlich gedankt für seine freundlichen und zuvorkommenden Informationen

²³⁰ Für seine Hinweise dankt der Verfasser dieser Arbeit Giuseppe G.A. Celano sehr herzlich.

3) Quellen- und Literaturverzeichnis

3.1) Gedruckte Quellen (Quelleneditionen)

Thomas *Wallnig*, Die Briefe von Johann Georg Eckhart an Bernhard Pez in Text und Kommentar (Staatsprüfungsarbeit am Institut für Österreichische Geschichtsforschung, Wien 2001).

Thomas *Wallnig*, Thomas *Stockinger*, Die gelehrte Korrespondenz der Brüder Pez. Text, Regesten, Kommentare, Bd 1: 1709–1715 (Quelleneditionen des Instituts für Österreichische Geschichtsforschung, Bd. 2/1, München/Wien 2010).

Thomas *Stockinger*, Thomas *Wallnig*, Patrick *Fiska*, Ines *Peper*, Manuela *Mayer*, Die gelehrte Korrespondenz der Brüder Pez. Text, Regesten, Kommentare, Bd. 2 (2 Halbbände): 1716–1718 (Quelleneditionen des Instituts für Österreichische Geschichtsforschung, Bd. 2/1 und Bd. 2/2, Wien 2015).

3.2) Sekundärliteratur

Anne *Bailiot* (Hg.), Briefe und Texte aus dem intellektuellen Berlin um 1800 (Berlin 2010–2015), online unter:
<https://www.literatur.hu-berlin.de/de/berliner-intellektuelle-1800-1830>
und <http://www.berliner-intellektuelle.eu/>; letzter Zugriff auf beide URLs: 30.05.2019.

David *Bamman*, Gregory *Crane*, The Design and Use of Latin Dependency Treebank. In: Jan *Hajič*, Joakim *Nivre* (Hgg.), Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (Prag 2006), 67–78; online unter: <http://ufal.mff.cuni.cz/tlt2006/pdf/110.pdf>; letzter Zugriff: 04.07.2019.

David *Bamman*, Gregory *Crane*, The Latin Dependency Treebank in a Cultural Heritage Digital Library. In: *Association for Computational Linguistics* (Hg.), Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data – LaTeCH 2007 (Prag 2007), 33–40; online unter: <https://www.aclweb.org/anthology/W07-0905>; letzter Zugriff: 04.07.2019.

David *Bamman*, Gregory *Crane*, Marco *Passarotti*, Savina *Raynaud*, A Collaborative Model of Treebank Development. In: Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (Bergen 2007), 1–6; online unter: https://www.researchgate.net/publication/28584823_A_Collaborative_Model_of_Treebank_Development; letzter Zugriff: 31.05.2019.

David *Bamman*, Gregory *Crane*, Marco *Passarotti*, Savina *Raynaud*, Guidelines for the Syntactic Annotation of Latin Treebanks, version 1.3. (Tufts/Medford 2007), 1–48; online unter: https://itreebank.marginalia.it/doc/2007_Passa+Bamman+Crane+Raynaud_Guidelines%20Tb.pdf; letzter Zugriff: 31.05.2019.

David *Bamman*, Roberto *Busa*, Gregory *Crane*, Marco *Passarotti*, The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin. In: Proceedings of the LREC 2008 (Marrakech 2008), 71–76; online unter: http://www.lrec-conf.org/proceedings/lrec2008/pdf/25_paper.pdf; letzter Zugriff: 31.05.2019.

David *Bamman*, Gregory *Crane*, Building a Dynamic Lexicon form a Digital Library. In: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (New York 2008), 11–20; online unter: <http://people.ischool.berkeley.edu/~dbamman/pubs/pdf/jcdl2008.pdf>; letzter Zugriff: 30.05.2019.

- David *Bamman*, Gregory *Crane*, Structured Knowledge for Low-Resource languages: The Latin and Greek Dependency Treebanks. In: Proceedings of the Text Mining Services 2009 (Leipzig 2009), 1–10; online unter: <https://pdfs.semanticscholar.org/2b86/9b06c81b19c2a76bcdcbd7dba5a6aad2f672.pdf>; letzter Zugriff: 04.07.2019.
- Piotr *Bański*, Susanne *Haaf*, Martin *Mueller*, Lightweight Grammatical Annotation in the TEI. New Perspectives. In: *European Language Resource Association* (Hg.), [Proceedings of the Eleventh International Conference on Language Resources and Evaluation \(LREC-2018\)](https://www.lrec-conf.org/proceedings/lrec2018/pdf/422.pdf) (Miyazaki 2018), 1795–1802; online unter: <http://www.lrec-conf.org/proceedings/lrec2018/pdf/422.pdf>; gesamter Tagungsband: <https://www.aclweb.org/anthology/L18-1>; letzter Zugriff auf beide URLs: 30.05.2019.
- Claudia *Bedini*, Núria Bertomeu *Castelló*, Dipanjan *Das*, Kuzman *Ganchev*, Yoav *Goldberg*, Keith *Hall*, Jungmee *Lee*, Ryan *McDonald*, Joakim *Nivre*, Yvonne *Quirnbach-Brundage*, Slav *Petrov*, Oscar *Täckström*, Hao *Zhang*, [Universal Dependency Annotation for Multilingual Parsing](https://ryanmcd.github.io/papers/treebanksACL2013.pdf). In: *Association for Computational Linguistics* (Hg.), Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Sofia 2013), 92–97; online unter: <https://ryanmcd.github.io/papers/treebanksACL2013.pdf>; letzter Zugriff: 02.08.2019.
- Samuel R. *Bowman*, Miriam *Connor*, Timothy *Dozat*, Christopher D. *Manning*, Marie-Catherine *de Marneffe*, Natalia *Silveira*, [More constructions, more genres: Extending Stanford Dependencies](https://www.aclweb.org/anthology/W13-3721). In: Proceedings of the Second International Conference on Dependency Linguistics: DepLing; Prague, August 27–30, 2013 (Prag 2013), 187–196; online unter: <https://www.aclweb.org/anthology/W13-3721>; letzter Zugriff: 02.08.2019.
- Noah *Bubenhof*, Marek *Konopka*, Roman *Schneider*, Präliminarien einer Korpusgrammatik (Tübingen 2014).

Peter *Burke*, Küchenlatein. Sprache und Umgangssprache in der Frühen Neuzeit (Kleine Kulturwissenschaftliche Bibliothek 14, Berlin 1989).

Peter *Burke*, Roy *Porter* (Hgg.), Language, Self, and Society (Oxford 1991).

Peter *Burke*, Languages and Communities in Early Modern Europe (Cambridge 2004).

Kai-Uwe *Carstensen*, Christian *Ebert*, Cornelia *Ebert*, Susanne *Jekat*, Ralf *Klabunde*, Hagen *Langer* (Hgg.), Computerlinguistik und Sprachtechnologie. Eine Einführung (Heidelberg ³2010).

Gregory *Crane*, Building a Digital Library. The Perseus Project as a Case Study in the Humanities. In: Proceedings of the first ACM International Conference on Digital libraries (New York 1996), 3–10; online unter: https://www.researchgate.net/publication/234791911_Building_a_digital_library_The_Perseus_Project_as_a_case_study_in_the_humanities; letzter Zugriff: 04.07.2019.

Gregory *Crane*, Cultural Heritage Digital Libraries. Needs and Components. In: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (London 2002), 626–663; online unter: <http://www.perseus.tufts.edu/~ababeu/ecdl2002.pdf>; letzter Zugriff: 04.07.2019.

Gregory *Crane*, Clifford *Wulfman* et al., Towards a Cultural Heritage Digital Library. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) (Washington 2003), 75–86; online unter: <http://www.ccs.neu.edu/home/dasmith/jcdl2003.pdf>; letzter Zugriff: 04.07.2019.

Ulrike Czeitschner, Thierry Declerck, Karlheinz Moerth, Claudia Resch, Gerhard Budin, A Text Technology Infrastructure for Annotating Corpora in the eHumanities. In: Francesca Borri, Stefan Gradmann, Carlo Meghini, Heiko Schuldt (Hgg.), Research and Advanced Technology for Digital Libraries. Proceedings of the International Conference on Theory and Practice of Digital Libraries (Berlin/Heidelberg 2001), 457–460.

Ulrike Czeitschner, Barbara Krautgartner, Claudia Resch, Eva Wohlfarter, Introducing the Austrian Baroque Corpus. Annotation and Application of a Thematic Research Collection. In: Florentina Armaselu, Marten During, Catherine Jones, René Leboutte, Lars Wieneke (Hgg.), Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age. Luxembourg, December 5-6, 2013 (Luxemburg 2016); online unter: <https://pdfs.semanticscholar.org/6728/4d7397f69e2636c9f9cf9cb4add28c80ba37.pdf>; letzter Zugriff: 02.08.2019.

Ulrike Czeitschner, Claudia Resch, Morphosyntaktische Annotation historischer deutscher Texte: Das Austrian Baroque Corpus. In: Wolfgang Dressler, Claudia Resch (Hgg.), Digitale Methoden der Korpusforschung in Österreich (Veröffentlichungen zur Linguistik und Kommunikationsforschung, Bd. 30, Wien 2017), 39–62.

Dipanjan Das, Slav Petrov, [Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections](#). In: *Association for Computational Linguistics* (Hg.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Portland/Oregon 2011), 600–609; online unter: <http://static.googleusercontent.com/media/research.google.com/sv//pubs/archive/37071.pdf>; letzter Zugriff: 05.08.2019.

- Dipanjan Das, Ryan McDonald, Slav Petrov, [A Universal Part-of-Speech Tagset](#). In: *European Language Resources Association ELRA* (Hg.), [Proceedings of the Eighth International Conference on Language Resources and Evaluation \(LREC-2012\)](#) (Istanbul 2012), 2089–2096; online unter: <http://www.petrovi.de/data/universal.pdf>; letzter Zugriff: 05.08.2019.
- Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, Klaus-Peter Wegera, HiTS. Ein Tagset für historische Sprachstufen des Deutschen. In: *Journal for Language Technology and Computational Linguistics – JLCL* 28/1 (2013), 85–137.
- Timothy Dozat, Filip Ginter, Katri Haverinen, Christopher D. Manning, Marie-Catherine de Marneffe, Joakim Nivre, Natalia Silveira, [Universal Stanford Dependencies: A cross-linguistic typology](#). In: *European Language Resources Association ELRA* (Hg.), [Proceedings of the Ninth International Conference on Language Resources and Evaluation \(LREC-2014\)](#) (Reykjavik 2014), 4585–4592; online unter: https://nlp.stanford.edu/pubs/USD_LREC14_paper_camera_ready.pdf; letzter Zugriff: 05.08.2019.
- Steffen Eger, Alexander Mehler, Tim von der Brück, Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods. In: *Association for Computational Linguistics, The Asian Federation of Natural Language Processing* (Hgg.), *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (Beijing 2015), 105–113; online unter: <http://www.aclweb.org/anthology/W15-3716>; letzter Zugriff: 31.05.2019.

- Steffen *Eger*, Rüdiger *Gleim*, Alexander *Mehler*, Lemmatization and morphological tagging in German and Latin. A comparison and a survey of the state-of-the-art. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (Portorož 2016), 1507–1513; <https://pdfs.semanticscholar.org/99b3/691ac04d7a196be1b7115189aca51f143c6d.pdf>; letzter Zugriff: 31.05.2019.
- Cornelia *Faustmann*, Gottfried *Glaßner*, Thomas *Wallnig* (Hgg.), Melk in der barocken Gelehrtenrepublik. Die Brüder Bernhard und Hieronymus Pez, ihre Forschungen und Netzwerke (Thesaurus Mellicensis, Bd. 2, Melk 2014).
- Joshua A. *Fishman*, The Sociology of Language. An Interdisciplinary Social Science Approach to Language in Society (Rowley/Mass. 1972).
- Tim *Geelhaar*, Rüdiger *Gleim*, Alexander *Mehler*, Tim *vor der Brück*, Towards a Network Model of the Coreness of Texts. An Experiment in Classifying Latin Texts Using the TTLab Latin Tagger. In: Chris *Biemann*, Alexander *Mehler* (Hgg.), Text Mining. From Ontology Learning to Automated Text Processing Applications (Berlin/New York 2015), 87–112; online unter: https://www.researchgate.net/publication/281461116_Towards_a_Network_Model_of_the_Coreness_of_Texts_An_Experiment_in_Classifying_Latin_Texts_Using_the_TTLab_Latin_Tagger; letzter Zugriff: 31.05.2019.
- André *Hagenbruch*, Flache Satzverarbeitung. In: Kai-Uwe *Carstensen*, Christian *Ebert*, Cornelia *Ebert*, Susanne *Jekat*, Ralf *Klabunde*, Hagen *Langer* (Hgg.), Computerlinguistik und Sprachtechnologie. Eine Einführung (Heidelberg ³2010), 264–279.
- Jan *Hajič*, Building a Syntactically Annotated Corpus. The Prague Dependency Treebank. In: Eva *Hajičová* (Hg.), Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová (Prag 1998), 12–19.

Michael *Hess*, Tagging (Zürich 2006),

online unter: <https://files.ifi.uzh.ch/cl/hess/classes/le/tag.0.1.pdf>;

letzter Zugriff: 30.05.2019.

Erhard *Hinrichs*, Tylman *Uhle*, Linguistische Annotation. In: Lothar *Lemnitzer*, Henning *Lobin* (Hgg.), Texttechnologie. Perspektiven und Anwendungen (Tübingen 2004), 217–243.

Fotis *Jannidis*, Perspektiven empirisch-quantitativer Methoden in der Literaturwissenschaft. Ein Essay. In: Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte 89/4 (2015), 657–661.

Fotis *Jannidis*, Hubertus *Kohle*, Malte *Rehbein* (Hgg.), Digital Humanities. Eine Einführung (Stuttgart 2017).

Fotis *Jannidis*, Grundlagen der Datenmodellierung. In: Fotis *Jannidis*, Hubertus *Kohle*, Malte *Rehbein* (Hgg.), Digital Humanities. Eine Einführung (Stuttgart 2017), 99–108.

Gard. B. *Jenset*, Barbara *McGillivray*, Quantitative Historical Linguistics (Oxford Studies in Diachronic and Historical Linguistics 26, Oxford 2017).

Holger *Keibel*, Marc *Kupietz*, Rainer *Perkuhn*, Korpuslinguistik (Paderborn 2012).

Adam *Kilgariff*, David *Tugwell*, Word Sketch. Extraction and Display of Significant Collocations for Lexikography. In: Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation (Toulouse 2001), 32–38; online unter: <https://www.kilgariff.co.uk/Publications/2001-KilgTugwell-ACLcollos-Sketches.pdf>; letzter Zugriff: 31.05.2019.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, David Tugwell, The Sketch Engine.
In: Sandra Vessier, Geoffrey Williams (Hgg.), Proceedings of the 11th
EURALEX International Congress (Lorient 2004), 105–115; online unter:
<https://euralex.org/publications/the-sketch-engine/>;

letzter Zugriff: 31.05.2019.

Adam Kilgarriff, Barbara McGillivray, Tools for Historical Corpus Research, and
a Corpus of Latin. In: Paul Bennett, Martin Durell, Silke Scheible, Richard
J. Whitt (Hgg.), New Methods in Historical Corpora (Korpuslinguistik und
interdisziplinäre Perspektiven auf Sprache – Corpus Linguistics and
Interdisciplinary Perspectives on Language Bd. 3, Tübingen 2013), 247–
256; online unter: [https://www.researchgate.net/publication/236857134_](https://www.researchgate.net/publication/236857134_Tools_for_historical_corpus_research_and_a_corpus_of_Latin)
[Tools_for_historical_corpus_research_and_a_corpus_of_Latin](https://www.researchgate.net/publication/236857134_Tools_for_historical_corpus_research_and_a_corpus_of_Latin);

letzter
Zugriff: 31.05.2019.

Ralf Klabunde, Automatentheorie und Formale Sprachen.

In: Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert,
Susanne Jekat, Ralf Klabunde, Hagen Langer (Hgg.), Computerlinguistik
und Sprachtechnologie. Eine Einführung (Heidelberg ³2010), 66–93.

Holger Kuße, Kulturwissenschaftliche Linguistik. Eine Einführung
(Göttingen 2012).

Sarah Lang, Review of Perseus Digital Library. In: *Institut für Dokumentologie
und Editorik e.V.* (Hg.), ride. A Review Journal for Digital Editions and
Resources (Köln 2018); online unter:

<https://ride.i-d-e.de/issues/issue-8/perseus/>; letzter Zugriff: 02.06.2019.

Hagen Langer, Syntax und Parsing. In: Kai-Uwe Carstensen, Christian Ebert,
Cornelia Ebert, Susanne Jekat, Ralf Klabunde, Hagen Langer (Hgg.),
Computerlinguistik und Sprachtechnologie. Eine Einführung (Heidel-
berg ³2010), 280–329.

Lothar Lemnitzer, Henning Lobin (Hgg.), Texttechnologie. Perspektiven und
Anwendungen (Tübingen 2004).

Angelika Linke, „Wer sprach warum wie zu einer bestimmten Zeit?“ Überlegungen zur Gretchenfrage der Historischen Soziolinguistik am Beispiel des Kommunikationsmusters ‚Scherzen‘ im 18. Jahrhundert. In: Ulrich Ammon, Jeroen Darquennes, Leigh Oakes, Sue Wright (Hgg.), *Sociolinguistica. Internationales Jahrbuch für europäische Soziolinguistik*, Bd. 13, Heft 1 (1999), 179–208.

Lothar Lemnitzer, Heike Zinsmeister, *Korpuslinguistik. Eine Einführung* (Tübingen 2015).

Susanne Lenz, *Korpuslinguistik* (Tübingen 2000).

Henning Lobin, *Computerlinguistik und Texttechnologie* (Paderborn 2010).

Bill MacCartney, Christopher D. Manning, Marie-Catherine de Marneffe, [Generating Typed Dependency Parses from Phrase Structure Parses](#). In: Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, Daniel Tapias, *European Language Resources Association ELRA* (Hgg.), [Proceedings of the Fifth International Conference on Language Resources and Evaluation \(LREC'06\)](#) (Genua 2006), 449–454; online unter: https://nlp.stanford.edu/pubs/LREC06_dependencies.pdf; letzter Zugriff: 05.08.2019.

Christopher D. Manning, Marie-Catherine de Marneffe, [The Stanford typed dependencies representation](#). In: *Association for Computational Linguistics* (Hg.), *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation* (Manchester 2008), 1–20; online unter: <https://nlp.stanford.edu/pubs/dependencies-coling08.pdf>; letzter Zugriff: 05.08.2019.

Klaus Mattheier, *Nationalsprachenentwicklung, Sprachenstandardisierung und Historische Soziolinguistik*. In: Ulrich Ammon, Klaus Mattheier, Peter Nelde (Hgg.), *Sociolinguistica. Internationales Jahrbuch für europäische Soziolinguistik*, Bd. 2, Heft 1 (1988), 1–9.

- Klaus *Mattheier*, Historische Soziolinguistik. Ein Forschungsansatz für eine künftige europäische Sprachgeschichte. In: Helga *Bister-Broosen* (Hg.), Beiträge zur historischen Stadtsprachenforschung (Schriften zur diachronen Sprachwissenschaft 8, Wien 1999), 223–234.
- Manuela *Mayer*, Das Chartular von St. Emmeran und seine Edition durch Bernhard *Pez*. In: *MIÖG* 125/2 (2017), 288–303.
- Ryan *McDonald*, Joakim *Nivre*, [Characterizing the Errors of Data-Driven Dependency Parsing Models](#). In: *Association for Computational Linguistics* (Hg.), Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (Prag 2007), 122–131; online unter: <https://www.aclweb.org/anthology/D07-1013>; letzter Zugriff: 05.08.2019.
- Barbara *McGillivray*, *Methods in Latin Computational Linguistics* (Bosten/Leiden 2014).
- Ilka *Mindt*, Methoden der Korpuslinguistik. Der korpusbasierte und der korpusgeleitete Ansatz. In: Iva *Kratochvílová*, Norbert Richard *Wolf* (Hgg.), *Kompodium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive* (Heidelberg 2010).
- Thomas *Müller*, Helmut *Schmid*, Hinrich *Schütze*, Efficient Higher-Order CRFs for Morphological Tagging. In: *Association for Computational Linguistics* (Hg.), Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (Seattle 2013), 322–332; online unter: <http://www.aclweb.org/anthology/D13-1032>; letzter Zugriff: 04.07.2019.
- Dietmar *Najock*, Helmut *Schmid*, Uwe *Springmann*, LatMor. A Latin Finite-State Morphology Encoding Vowel Quantity. In: Sabine *Erhart* (Hg.), *Open Linguistics* 2/1 (2016), 386–392; online unter: <https://www.degruyter.com/downloadpdf/j/opli.2016.2.issue-1/opli-2016-0019/opli-2016-0019.pdf>; letzter Zugriff: 04.07.2019.

- Sven *Naumann*, Hagen *Langer*, Parsing. Eine Einführung in die maschinelle Analyse natürlicher Sprache (Trier 1994), online unter: <https://www.uni-trier.de/fileadmin/fb2/LDV/Naumann/BUCH1-1.pdf>; letzter Zugriff: 30.05.2019.
- Gerhard *Nickel*, Einführung in die Linguistik. Entwicklung, Probleme, Methoden (Berlin 1979).
- Heidrun *Pelz*, Linguistik. Eine Einführung (Hamburg ⁴1999).
- Marco *Passarotti*, Verso il Lessico Tomistico Biculturale. La treebank dell Index Thomisticus. In: Diego *Femia*, Raffaella *Petrilli* (Hgg.), Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, 14–16 Settembre 2006 (Roma 2007), 187–205.
- Marco *Passarotti*, Leaving Behind the Less-Resourced Status. The Case of Latin through the Experience of the Index Thomisticus Treebank. In: Proceedings of the 7th SaLTMiL Workshop on the Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010 (La Valletta 2010), 27–32; online unter: <http://docplayer.net/7749747-Leaving-behind-the-less-resourced-status-the-case-of-latin-through-the-experience-of-the-index-thomisticus-treebank.html>; letzter Zugriff: 31.05.2019.
- Michael *Piotrowski*, Digital Humanities. An Explication. In: Manuel *Burghardt*, Claudia *Müller-Birn* (Hgg.), INF-DH-2018. Workshopband 25. Sept. 2018 (Berlin 2018); online unter: <https://dl.gi.de/handle/20.500.12116/17004>; letzter Zugriff: 03.08.2019.
- Bodo *Plachta*, Editionswissenschaft. Eine Einführung in Methode und Praxis der Edition neuerer Texte (Stuttgart/Ditzinger ³2013).

- Andrea *Rapp*, Manuelle und automatische Annotation. In: Fotis *Jannidis*, Hubertus *Kohle*, Malte *Rehbein* (Hgg.), Digital Humanities. Eine Einführung (Stuttgart 2017), 253–267.
- Georg *Rehm*, Texttechnologische Grundlagen. In: Kai-Uwe *Carstensen*, Christian *Ebert*, Cornelia *Ebert*, Susanne *Jekat*, Ralf *Klabunde*, Hagen *Langer* (Hgg.), Computerlinguistik und Sprachtechnologie. Eine Einführung (Heidelberg 32010), 159–168.
- Claudia *Resch*, „Etwas für alle.“ Ausgewählte Texte von und mit Abraham a Sancta Clara digital. In: Zeitschrift für digitale Geisteswissenschaften 2 (2017); online unter:
http://www.zfdg.de/sites/default/files/pdf/santa_clara_2015.pdf;
 letzter Zugriff: 05.08.2019.
- Claudia *Resch*, Linguistisch annotierte historische Texte stilistisch auswerten. Musterhaft vorkommende Wortverbindungen im Austrian Baroque Corpus. In: Lisa *Dücker*, Stefan *Hartmann*, Renata *Szczepaniak* (Hgg.), Historische Korpuslinguistik [Jahrbuch für Germanistische Sprachgeschichte Bd. 10, Berlin/Boston 2019 (in Vorb.)].
- Philip *Resnik*, Daniel *Zeman*, [Cross-Language Parser Adaptation between Related Languages](#). In: *Asian Federation of Natural Language Processing* (Hg.), Proceedings of IJCNLP 2008 Workshop on NLP for Less Privileged Languages (Hyderabad 2008), 35–42; online unter:
<http://ufal.mff.cuni.cz/~zeman/publikace/2008-01/padapt-hyderabad-05c-postfinal.pdf>; letzter Zugriff: 05.08.2019.

- Paul T. *Roberge*, *Language History and Historical Sociolinguistics. Sprachgeschichte und historische Soziolinguistik*.
In: Ulrich *Ammon*, Norbert *Dittmar*, Klaus *Mattheier*, Peter *Trudgill* (Hgg.), *Sociolinguistics/Soziolinguistik. An International Handbook of the Science of Language and Society/Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science, Bd. 3, 3. Teilband., Berlin ²2006)*, 2306–2315.
- Patrick *Sahle*, *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels* (3 Bde., Schriften des Instituts für Dokumentologie und Editorik 7–9, Norderstedt 2013).
- Patrick *Sahle*, *Digitale Editionen*. In: Fotis *Jannidis*, Hubertus *Kohle*, Malte *Rehbein* (Hgg.), *Digital Humanities. Eine Einführung* (Stuttgart 2017), 234–249.
- Patrick *Sahle*, Georg *Vogeler*, *XML*. In: Fotis *Jannidis*, Hubertus *Kohle*, Malte *Rehbein* (Hgg.), *Digital Humanities. Eine Einführung* (Stuttgart 2017), 128–146.
- Felix *Sasaki*, Andreas *Witt*, *Linguistische Korpora*. In: Lothar *Lemnitzer*, Henning *Lobin* (Hgg.), *Texttechnologie. Perspektiven und Anwendungen* (Tübingen 2004), 195–216.
- Carmen *Scherer*, *Korpuslinguistik* (Heidelberg 2006).
- Peter *Schlobinski*, *Grundfragen der Sprachwissenschaft. Eine Einführung in die Welt der Sprache(n)* (Göttingen 2014).
- Wolfgang *Schmale* (Hg.), *Digital Humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität* (Stuttgart 2015).

Helmut *Schmid*, SFST Manual, o.O, o.D., online unter: <https://www.cis.uni-muenchen.de/~schmid/tools/SFST/data/SFST-Manual.pdf>; letzter Zugriff: 04.07.2019.

Helmut *Schmid*, [Probabilistic Part-of-Speech Tagging Using Decision Trees](#). In: Proceedings of the International Conference on New Methods in Language Processing, (Manchester 1994), 44–49; online unter: <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf>; letzter Zugriff: 05.08.2019.

Helmut *Schmid*, Parsing I. Seminarskript (Stuttgart 2003), online unter: <http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schmid/ParsingI/parsing.pdf>; letzter Zugriff: 04.07.2019.

Tobias *Schnabel*, Hinrich *Schütze*, FLORS: Fast and Simple Domain Adaption for Part-of-Speech Tagging. In: *Association for Computational Linguistics*, Sharon *Goldwater* (Hgg.), Transactions of the Association for Computational Linguistics 2 (2014), 15–26; online unter: <http://acl2014.org/acl2014/Q14/pdf/Q14-1005.pdf>; letzter Zugriff: 31.05.2019.

Manfred *Thaller*, Geschichte der Digital Humanities. In: Fotis *Jannidis*, Hubertus *Kohle*, Malte *Rehbein* (Hgg.), Digital Humanities. Eine Einführung (Stuttgart 2017), 3–12.

Manfred *Thaller*, Digital Humanities als Wissenschaft. In: Fotis *Jannidis*, Hubertus *Kohle*, Malte *Rehbein* (Hgg.), Digital Humanities. Eine Einführung (Stuttgart 2017), 13–18.

Yoshimasa *Tsuruoka*, Yusuke *Miyao*, Jun'ichi *Kazama*, Learning with Lookahead. Can History-Based Models Rival Globally Optimized Models? In: *Association for Computational Linguistics* (Hg.), Proceedings of the Fifteenth Conference on Computational Natural Language Learning (Portland 2011), 238–246; online unter: <http://www.aclweb.org/anthology/W11-0328>; letzter Zugriff: 31.05.2019.

Wim *Vandenbussche*, Roland *Willemyns*: Historical Sociolinguistics. Coming of Age? In: Ulrich *Ammon*, Klaus *Mattheier*, Peter *Nelde* (Hgg.), *Sociolinguistica. Internationales Jahrbuch für Europäische Soziolinguistik*, Bd. 20 (2006), 146–165.

Heinz *Vater*, Einführung in die Sprachwissenschaft (München ²1994).

Thomas *Walach*, Geschichte des virtuellen Denkens (Wiesbaden 2018).

Françoise *Waquet*, Le Latin ou l'Empire d'un Signe. XVIe – XXe Siècle (Paris 1999).

Thomas *Wallnig*, Gasthaus und Gelehrsamkeit. Studien zu Herkunft und Bildungsweg von Bernhard Pez OSB vor 1709 (Veröffentlichungen des Instituts für Österreichische Geschichtsforschung 48, Wien 2007; zugl.: Dissertation, Univ. Graz 2004).

Thomas *Wallnig*, Gelehrtenkorrespondenzen und Gelehrtenbriefe. In: Josef *Pauser*, Martin *Scheutz*, Thomas *Winkelbauer* (Hgg.), *Quellenkunde der Habsburgermonarchie (16.–18. Jahrhundert). Ein exemplarisches Handbuch* (Mitteilungen des Instituts für Österreichische Geschichtsforschung, Ergänzungsband 44, Wien/München 2004), 813–827.

Thomas *Wallnig*, Critical Monks. The German Benedictines, 1680–1740 (Scientific and Learned Cultures and Their Institutions, Bd. 25, Leiden/Boston 2019).

- Norbert Richard *Wolf*, Korpora in der Korpuslinguistik. In:
Iva *Kratochvílová*, Norbert Richard *Wolf* (Hgg.),
Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-
tschechischer Perspektive (Heidelberg 2010), 17–25.
- Daniel *Zeman*, [Reusable Tagset Conversion Using Tagset Drivers](http://lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf). In: *European Language Resources Association ELRA* (Hg.), Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May–1 June 2008 (Marrakech 2008), 213–218; online unter: http://lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf; letzter Zugriff: 05.08.2019,
- Heike *Zinsmeister*, Korpora. In: Kai-Uwe *Carstensen*, Christian *Ebert*, Cornelia *Ebert*, Susanne *Jekat*, Ralf *Klabunde*, Hagen *Langer* (Hgg.), Computerlinguistik und Sprachtechnologie. Eine Einführung (Heidelberg 2010), 482–491.

3.3) Wörterbücher und Grammatiken

- Charles *Du Fresne Du Cange*, Lorenz *Diefenbach*, *Lexicon mediae et infimae latinitatis* (Darmstadt 1997; unveränd. reprogr. Nachdr. d. Ausg. Frankfurt a.M. 1857).
- Der Neue Georges. Ausführliches lateinisch-deutsches Handwörterbuch (2 Bde.), ursprüngl. ausgearb. v. Karl-Ernst *Georges*, bearb. v. Heinrich *Georges*, neu hrsg. v. Thomas *Baier* & bearb. v. Tobias *Dänzer* (Darmstadt 2013).
- Hermann *Menge*, Lehrbuch der lateinischen Syntax und Semantik, bearb. v. Thorsten *Burkhard*, Markus *Schauer* (5. durchges. u. verb. Aufl. Darmstadt 2012).

3.4) Ausgewählte Online-Ressourcen

Corpus Thomisticum:

<http://www.corpusthomisticum.org/>

<http://www.corpusthomisticum.org/it/index.age>

e-Pisolarium: <http://ckcc.huygens.knaw.nl/>

Index Thomisticus Treebank:

<https://itreebank.marginalia.it/view/ittb.php>

<https://itreebank.marginalia.it/view/resources.php>

https://github.com/UniversalDependencies/UD_Latin-ITTB/tree/master

Intratext: <http://www.intratext.com>

Lacus Curtius:

<http://penelope.uchicago.edu/Thayer/E/Roman/home.html>

LiLA: <https://lila-erc.eu/about/>

LatMor:

<http://www.cis.uni-muenchen.de/~schmid/tools/LatMor/>

<http://www.cis.uni-muenchen.de/~schmid/tools/LatMor/LatMor-tag-list.txt>

<http://www.cis.uni-muenchen.de/~schmid/tools/SFST/>

<http://www.cis.uni-muenchen.de/~schmid/tools/SFST/data/SFST-Manual.pdf>

MarMot: <http://cistern.cis.lmu.de/marmot/>

Musisque Deoque:

<http://www.mqdq.it>

Perseids – Arethusa (Treebanking Tool):

<http://www.perseids.org/tools/arethusa/app/>

<http://sosol.perseids.org/sosol/signin>

<http://www.perseids.org/apps/treebank>

Perseus Digital Library (altes Interface):

<http://www.perseus.tufts.edu/hopper/>

Perseus Digital Library (neues Interface):

<https://scaife.perseus.org/>

Perseus Digital Library – Greek & Latin Treebanks:

https://perseusdl.github.io/treebank_data/

Pez-Nachlass digital:

<https://unidam.univie.ac.at/nachlass/195>

Prague Dependency Treebank:

<https://ufal.mff.cuni.cz/pdt3.5>

Sketch Engine:

<https://www.sketchengine.eu/>

<https://www.sketchengine.eu/latin-part-of-speech-tagset/>

Stuttgart-Tübingen-Tagset:

<http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1995.pdf>

<http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>

TEI: <https://tei-c.org/>

TokenEditor:

<https://clarin.oeaw.ac.at/tokenEditor/>

<https://www.oeaw.ac.at/acdh/tools/tokeneditor/>

Tree Tagger:

<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Latin-parameter-file-readme>

<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

Universal Dependencies:

<http://universaldependencies.org/introduction.html>

<http://universaldependencies.org/treebanks/la-comparison.html>

Wien(n)erisches Diarium:

<https://www.oeaw.ac.at/acdh/about/news-archive/news-detail/article/schlagzeilen-von-anno-dazumals-wandern-ins-web-1/>

<https://www.oeaw.ac.at/acdh/projects/wienerisches-diarium-digital/>

4) Anhang 1: Liste ausgewählter maschineller Annotationsfehler

Folgende Liste versteht sich weder als vollständig noch als sonderlich systematisch oder repräsentativ. Es soll durch sie nicht auf die Quantität maschineller Annotationsfehler aufmerksam gemacht werden. Vielmehr möchte sie eine Auswahl dessen bieten, welche *Arten* von Fehlern entweder einmalig oder *tendenziell* wiederholt aufgetreten sind.

1. Zuweilen Verwechslung von Nom. und Akk. Pl. der Konsonanten- und der Mischstämme (enden beide auf -es, z.B. *codices*).
2. Zuweilen Verwechslung von Gen. Sg. und Nom. Pl. der O-Stämme (enden beide auf -i).
3. Zuweilen Verwechslung von Nom. und Akk. Sg. bei Neutra der O-Stämme (beide Casus enden auf -um).
4. Zuweilen Verwechslung von Gen. Sg., Dat. Sg. und Nom. Pl. der A-Stämme (enden alle auf -ae).
5. *ipsi*: Nom. Pl. m. und Dat. Sg. wurden häufig verwechselt.
6. Gen. Pl. von *hic/haec/hoc*: *horum* wurde mitunter auf *hora* zurückgeführt.
7. Die Präposition *ad* wurde mitunter als ADV getaggt.
8. Adverbia wurden grundsätzlich *immer* mit einem Steigerungsgrad (Positiv, Comparativ oder Superlativ) versehen, auch wenn es dgl. nicht gibt (etwa bei Adverbien wie *praesertim*, *raptim*, *statim*, *tamen*, ...).
9. Konjunktivformen der a-Stämme wurden regelmäßig falsch lemmatisiert (z.B. *amem* > Lemma: *amis*).
10. Gerundium und Gerundiv wurden häufig verwechselt.
11. Das Perfekt von *ferre* (Stamm *tul-*) wurde regelmäßig falsch lemmatisiert (einschließlich Composita).
12. Generell traten Probleme bei der Lemmatisierung unregelmäßig flektierender Verben auf (z.B. bei *velle*, *malle*, *posse*; manchmal bei *esse*: *eras* > zuweilen lemmatisiert mit *era*).
13. Konjunktivformen von *velle* (z.B. *velim*) wurden so gut wie *immer* mit einem falschem Lemma versehen; so wurde *velis* etwa mit *velus* lemmatisiert und *velim* mit *velis*, beides bei korrekter Morphologie.

14. Kontrahierte Verbalformen bereiteten mitunter Probleme: *fecere* wurde mit *feco* lemmatisiert (bei korrekter Morphologie), bei *approbarunt* fehlte die Morphologie vollständig.
15. Flektierte Formen von *putare* wurden oft mit *putus* lemmatisiert.
16. *constitui*: lemmatisiert mit *constio*.
17. *possem*: lemmatisiert mit *possis* (bei korrekter Morphologie).
18. *scripseras* lemmatisiert mit *scripsera*, getaggt als NN Acc. Pl. fem.
19. *prodibit*: lemmatisiert mit *prodibo* statt mit *prodeo*.
20. *impendam*: lemmatisiert mit *impo*.
21. *imposui*: lemmatisiert mit *imposus*.
22. *exposui* lemmatisiert mit *exponeo* (bei korrekter Morphologie).
23. *risit*: lemmatisiert mit *ritto*.
24. *edet* (von *edere*): lemmatisiert mit *edeo*.
25. *excipiebar*: lemmatisiert mit *excipiebaus* (sic!).
26. *solvam*: lemmatisiert mit *solvus* (bei korrekter Morphologie).
27. *mones*: lemmatisiert mit *mo* (sic!).
28. *nominem*: lemmatisiert mit *nomo* (bei korrekter Morphologie).
29. *Jesuitae*: lemmatisiert mit *Jesubo* (sic!).
30. *scis*: lemmatisiert mit *scus* (bei korrekter Morphologie).
31. *ingentibus*: lemmatisiert mit *ingo* (bei korrekter Morphologie).
32. *scribe*: lemmatisiert mit *scrio* (bei korrekter Morphologie).
33. *scivissem*: lemmatisiert mit *scivisso* (bei korrekter Morphologie).
34. *emi* (von *emere*): lemmatisiert mit *eo*.
35. *ades* (von *adesse*): lemmatisiert mit *ado*.
36. *reperisti*: lemmatisiert mit *reperistus* (bei korrekter Morphologie).
37. *sancte*: fallweise lemmatisiert mit *sano*.
38. *dico*: fallweise lemmatisiert mit *do*.
39. *ultimam*: lemmatisiert mit *ultrus*.
40. *insigne*: lemmatisiert mit *inso*.
41. *favi* (von *faveo*) lemmatisiert mit *fo* (bei korrekter Morphologie).
42. *erunt*: lemmatisiert mit *esum* (bei korrekter Morphologie).
43. *ii* (PDEM): lemmatisiert mit *ium*.
44. *ADV clare* (von *clarus*): lemmatisiert mit *clo* (sic!).

5) Anhang 2/1: Belegstellen für semant. Analysen 1 (aus Briefen Eckharts an die Brüder Pez)

BE 1 (Johann Georg Eckhart an Bernhard Pez; 1717-12-30;
Hannover; Edition 2015: Brief 870²³¹):

Codex, quem desideras, non est in regia bibliotheca, quam Hanoverae ego curo, sed Guelfebyti. Curabo tamen, ut quam primum ad me et inde sumtibus regiis per dominum de Huldensberg ad te veniat, certus de caetero illaesum rediturum. Eleganter scriptus est in forma, quam **octavam** vocant.

BE 2 (wie BE 1):

Ego de caetero quondam apud illustrissimum comitem de Beichling tunc primarium Augusti Poloniarum regis status ministrum vidi inter magnam, quam ille collegerat sumtu in re nihili stupendo, **librorum** magicorum copiam **codicem** seculo decimoquarto scriptum et Bennonis episcopi Magia inscriptum.

BE 3 (Johann Georg Eckhart an Bernhard Pez; 1718-01-27; Hannover;
Edition 2015: Brief 891):
Luneburgo reversus Guelfenbyti recepi **codicem**, cui inserta
Bennonis episcopi opuscula gemina.

BE 4 (wie BE 3):

Codice hoc Guelfenbytani non adeo diu carere possunt.

BE 5 (wie BE 3):

Est in bibliotheca augusta **codex**, qui continet Eschenbachi poetae Germanici Germanicum opus De occupatione Terrae Sanctae per Godefridum Bullionensem.

BE 6 (Johann Georg Eckhart an Bernhard Pez; 1718-06-26; Hannover;
Edition 2015: Brief 955):

Age vero, et nobis brevi ipsam **codicis** Zwetlensis editionem profer.

²³¹ Ein- bis dreistellige Nummerierungen der Briefe entsprechen jenen Nummerierungen in den Editionen der Jahre 2010/2015 (vgl. Anm. 15).

BE 7 (wie BE 6):

Mereretur iterum edi et cum manuscriptis conferri **Codex** epistolarum Carolinus.

BE 8 (wie BE 6):

Regia mea bibliotheca habet exemplar Flacii manu ex vetustissimo **codice** transscritum.

BE 9 (wie BE 6):

Codicem Theotiscum Herbipolensem tam solícite custodiri, ut communicari non possit, doleo.

BE 10 (Johann Georg Eckhart an Hieronymus Pez; 1718-12-25; Hannover; Edition 2015: Brief 1031):

Codex etiam epistolarum Rudolphi I. imperatoris, quem Czerwenka habuit et pene integrum edidit, huc pertinet.

BE 11 (wie BE 10):

Aliqua Leibnitius in **Codice** diplomatico profert.

BE 12 (Johann Georg Eckhart an Bernhard Pez; 1718-12-25; Hannover; Edition 2015: Brief 1032):

Recepi nuper accuratissimam collationem Otfridi Paraphraseos Evangeliorum, quam olim cum Vaticano bibliothecae Palatinae **codice** instituit Fridericus Rostgardus Danus.

BE 13 (wie BE 12):

Veniet etiam brevi ad me **Codex** Otfridinus ex Viennensi descriptus cum versione Latina Dieterici Stadenii τοῦ νῦν ἐν ἁγίοις et amplissimo eiusdem commentario.

BE 14 (wie BE 12):

Nuper amicus misit **codicem membranaceum**, qui continet Hermanni
Zoest de Monasterio, professi in monasterio de Campo S. Mariae
Cisterciensis ordinis dioeceseos Monasteriensis, opuscula De ecclesiastica
potestate et papali ac De vocibus diffinitivis in conciliis generalibus.

BE 15 & 16 (Johann Georg Eckhart an Bernhard Pez; 1719-09-03;

Hannover; Edition 2001: Brief GE 12²³²):

Est Passaviae, in bibliotheca an archivio nescio, inter vetustos **codices**
codex capitularium legumque veterum, quem conferri auf videre velim;
an tibi ibi amici?

BE 17 (Johann Georg Eckhart an Bernhard Pez; 1720-11-10; Münster;

Edition 2001: Brief GE 15):

Li enim, qui Francofurti aurum in tenuissima folia cudunt, omnes pene
e Westphalia **codices membranaceos** avexerunt et nuper admodum
Teclenburgi duodecim **volumina in folio** nacti sunt, quae quovis
pretio redemissent.

BE 18 (Johann Georg Eckhart an Bernhard Pez; 1720-12-28;

Hannover; Edition 2001: Brief GE 16):

Desidero quidem **codicem** illum Theotiscum.

BE 19 & 20 (J. G. Eckhart an Bernhard Pez; 1721-01-01; Hannover;

Edition 2001: Brief GE 17):

Velim mecum is literas commutet, ubi Herbipoli erit, et citissime
perscribat, quos **codices**, quae cartularia, leges veteres, chronica,
codices Theotiscos et horum similia invenerit.

²³² Die Nummerierung der Briefe mit den Siglen „GE [Ziffer]“ entspricht der Nummerierung in der Edition des Jahres 2001 (vgl. Anm. 15). Es betrifft dies Briefe, die nach dem Jahr 1718 verfasst wurden und in den Editionen von 2010/15 noch nicht berücksichtigt werden konnten.

- BE 21** (Johann Georg Eckhart an Bernhard Pez; 1721-04-18; Hannover;
Edition 2001: Brief GE 20):
Abii enim medio mense Martio Hildesiam ibique vidi monasteria sancti
Michaelis et sancti Godehardi ordinis vestri, in hoc reperi **codicem**
vetustum Chronicon Reginonis continent[em], ex quo non pauca in
editione Pistorii correxi.
- BE 22** (wie BE 21):
Abbas, vir humanissimus, dicebat se viginti et aliquot veteres **codices**
habere.
- BE 23** (wie BE 21):
In bibliotheca abbatae **codex** vetustissimus et eleganter scriptus mihi
placebat, Sulpitii Severi Vitam Martini et alia de sancto Martino continens
scriptusque iussu domni Fredegisi.
- BE 24** (Johann Georg Eckhart an Bernhard Pez; 1721-09-19; Hannover;
Edition 2001: Brief GE 24):
Emi 300 imperialibus exemplar Evangeliorum Theotiscorum Otfridi
monachi Weissenburgensis, descriptum ex **codice** caesareo, et in
Latinam linguam translatum a Dieterico Stadenio, qui et omnium
vorum glossarium addidit, multa eruditione refertum.
- BE 25** (Johann Georg Eckhart an Bernhard Pez; 1722-02-26; Hannover;
Edition 2001: Brief GE 26):
Ubi historicae disquisitiones mihi aliquam temporis inter capedinem
concesserint, iam manus admovebo editioni Otfridi exactissimae ad
Vaticanum et Viennensem **codicem** revisae.
- BE 26** (Johann Georg Eckhart an Bernhard Pez; 1723-01-10; Hannover;
Edition 2001: Brief GE 29):
Sed amice, an graviter feres **Codicem** Udalrici inter eos comparere?

- BE 27** (Johann Georg Eckhart an Bernhard Pez; 1723-10-03; Hannover;
Edition 2001: Brief GE 30):
Visbecae etiam vetusta plura diplomata vidi et Cassellis copiam veterum
codicum manibus sociorum sancti Bonifacii scriptorum admiratus sum.
- BE 28** (Johann Georg Eckhart an Bernhard Pez; 1725-05-16; Hannover;
Edition 2001: Brief GE 31):
Perlustro iam **codices** bibliothecae cathedralis venerandos
constituique excerpta scitu digna ex iis edere.
- BE 29** (wie BE 28):
Nostine epistola[m] Timothei discipuli Pauli de Sanctimonio?
Extat ille in **codice** seculi VIII., sed suppositic[ius] videtur.
- BE 30** (wie BE 28):
Detexi **codicem** epistolarum Pauli cum glossa hibernica, quo sine dubio
sanctus Kilianus usus est, et homilias, quae sancti Burchardi fuere.
- BE 31** (wie BE 6):
Ego eiusdem materiae spissum **in octavo librum** in **membrana** seculo
XII. aut XIII. incipiente scriptum habeo.
- BE 32** (Johann Georg Eckhart an Bernhard Pez; 1719-05-28; Hannover;
Edition 2001: Brief: GE 11):
Mitto tamen Sermonem beati Volcuini ex veteri **membrana** descriptum.
- BE 33** (wie BE 21):
Unus in **membrana** purpurea literis Romanis aureis scriptus erat a
Samuehele presbytero.
- BE 34** (wie BE 26):
Per integrum fere annum nunc huc nunc illuc vagatus excussi
monasteriorum nostrorum **chartas** pulvere obsitas.

- BE 35** (Johann Georg Eckhart an Bernhard Pez; 1718-08-18; Hannover;
Edition 2015: Brief 981):
Iter meum ad vos non nisi liberiore animo suscipiam, hoc est finitis sex
tomis (**in folio**) prioribus Historiae serenissimae domus Brunsvicensis.
- BE 36** (wie BE 24):
[...] Bonaventura Tillesson , qui de rebus Suessionensibus quinque **in folio**
volumina scripsit nondum edita.
- BE 37** (Johann Georg Eckhart an Bernhard Pez; 1722-06-12; Hannover;
Edition 2001: Brief GE 27):
Imprimuntur iam et a me duo **in folio** tomi historicorum medii
aevi hactenus ineditorum.
- BE 38** (für eine Form von *liber*; siehe BE 2).
- BE 39** (wie BE 1):
Si in bibliothecis vestris glossaria vetera Theotisca, **libri** et poemata
idiomate vetusto scripta et alia huc pertinentia habeantur, eorum notitia
mihi erit gratissima.
- BE 40** (wie BE 3):
Lambecius eius **libro** secundo Bibliothecae meminit.
- BE 41**: siehe BE 31.
- BE 42** (wie BE 35):
Sed gemini huius nominis bibliopolae Lipsiae degunt. Johannis Friderici
filius est unus; alter Johannes Ludovicus est Gleditschius, apud quem
libri in Italia impressi extant.
- BE 43** (wie BE 35):
Expono inter caetera monumentum Celticum Parisiis repertum. Sed iam
exactiorem delineationem nactus sum a teste oculato et antiquitatum
perito, quam aeri incisam adiungo. **Librum** a me cum Radberto accipies.

BE 44 (wie BE 10):

Multa etiam acta publica Austriaca insunt Rymeri Actis Anglicanis,
rarissimo **libro**, quae tibi lubenter excerpti curabo.

BE 45 (wie BE 10):

Ita vero Murensis monasterii fundatio, Vita sanctae Odiliae et, quaecunque
Vignierius Oratorii presbyter in Gallico Des familles d'Alsace **libro** inter
probationes evulgavit, ad te pertinent.

BE 46 (Johann Georg Eckhart an Bernhard Pez; 1720-01-18; Hannover;

Edition 2001: Brief GE 13):

Ego primos fructus Svecicae pacis recepi, cistam nempe **libris** rarissimis
in Svecia editis refertam, quae mihi 361 imperialibus constat.

BE 47 (wie BE 17):

Misit me eo rex meus, ut rariora ex auctione **librorum** Mallincrotranorum
pro nostra bibliotheca coemerem. Et emi inde quadrigenta, et quod
superat, **volumina**, quae pretio sexcentorum imperialium venere.

BE 48 (wie BE 17):

Observavi tamen ex vestris patres Iburgenses, qui scilicet nobiscum
saepius conversantur ob principem episcopi sedem, quae Iburgi est,
[u]nos quosdam **libros** emisse, unde studia adhuc aliquatenus ipsis
curae esse conieci.

BE 49 (wie BE 19):

Integer ad te **liber** scribendus esset, si omnes meas hac in re
disquisitiones tibi aperire vellem.

BE 50 (Johann Georg Eckhart an Bernhard Pez; 1721-07-06; Hannover;

Edition 2001: Brief GE 22):

Non erat admodum eruditus, sed eruditionem amabat et alearum
chartarumque lusoriarum loco monachis suis **libros** in manus tradiderat.

BE 51 (wie BE 24):

Quis est Schifferus, quem dicunt de genealogiis nobilium
Bavariae scripsisse? Non novi illum aut eius **librum**.

BE 52 (wie BE 1):

Characteribus magicis et artibus prohibitis refertus erat, sancto viro ab
impostore dubio procul suppositus, uti alius de necromantia **libellus**
Johannae papissae adscriptus [...].

BE 53 (wie BE 6):

Magister Galterus an is est, qui Alexandreida scripsit? Eiusdem ineditum
De amore **libellum** Bremae amplissimus Maastrichtius servat.

BE 54 (wie BE 46):

Suscepi in me directionem et curam operis ingentis, cui titulus erit
Thesauri antiquitatum Germanicarum. In eo imprimentur omnes
libelli, qui antiquitates Germanicas illustrant.

BE 55 (wie BE 17):

Vidi multa apud ipsum iconia quibus ad nummos cudendos usus est
anabaptistarum olim rex Knipperdollingus. Innotuit etiam hic mihi
reverendus pater Strunck societatis Jesu, qui anonymus edidit
Westphaliā sanctam, **libellum** non inelegantem.

BE 56 (wie BE 28):

Constitui igitur edere **libellum** de numis Wirceburgensibus [...].

BE 57 (Johann Georg Eckhart an Bernhard Pez; 1728-01-13; Würzburg;

Edition 2001: Brief GE 33):

De Anonymo Ravennate non erravi; quanquam manuscripta eius non adeo
vetera supersint, certum tamen eum **libellum** seculo VII. compositum esse;
ut singula bene expensa produnt.

BE 58 (wie BE 57):

Meinbergii sive patris Seyfridi societatis Jesu **Libellus** sive Satyra
in Schannatum Fuldae forte per carnificem comburetur.

BE 59 (Johann Georg Eckhart an Bernhard Pez; 1719-01-12; Hannover;

Edition 2001: Brief GE 9):

Sed a serenissimo episcopo, ut verum fatear, ferme coactus,
quicquid hoc est, in **chartam** conieci.

BE 60 (wie BE 19):

Sed in ista temporum obscuritate quidam non nisi per coniecturas probari
possunt, ubi mallet authorum et **chartarum** verba producere.

BE 61 (Johann Georg Eckhart an Bernhard Pez; 1721-01-17; Hannover;

Edition 2001: Brief GE 18):

Sed iterum magna **chartarum** copia excutienda earumque falsitas
declaranda est.

BE 62 (wie BE 21):

Rosenthalius consiliarius regiminis ostendebat viginti **chartas**
geographicas [...].

BE 63 (wie BE 50):

Duo exemplaria operis tui recepi, sed in **charta** vulgari [...].

BE 64: siehe BE 50.

BE 65 (wie BE 24):

Splendidissime opus imprimi curabo, figuris aeneis ornatum et
in **charta** nitidissima.

BE 66 (Johann Georg Eckhart an Bernhard Pez; 1721-12-05; Hannover;

Edition 2001: Brief GE 25):

De Henselero nondum cogitare potui et, quae in **chartam** primo impetu
conieceram, iam, ut mihi non congruo, displicent [...].

BE 67: siehe BE 34.

BE 68 (wie BE 6):

Habemus etiam spissum **volumen** epistolarum Berengarii haeretici.

BE 69 (wie BE 46):

Eruditus quidam ecclesiae apud Gluckstadiensis minister, Siblingii nomine,
edit iam **volumen** scriptorum rerum Danicarum, quibus multa scitu digna
inseruit.

BE 70: siehe BE 47.

BE 71: siehe BE 17.

BE 72: siehe BE 36.

BE 73 (Johann Georg Eckhart an Bernhard Pez; 1722-06-24; Hannover;

Edition 2001: Brief GE 28):

Schannatus suadet, ut et diplomatum **volumen** edam.

6) Anhang 2/2: Belegstellen für semant. Analysen 2 (aus Briefen der Brüder Pez)

BP 1 (Bernhard Pez an René Massuet; 1711-01-25; Melk; Edition 2010: Brief 143):

Lambacense praeter catalogum abbatum dedit etiam **libellum**, quo Vitae beati Adalberonis episcopi Herbipolensis fundatoris Lambacensis, beati Altmanni episcopi Pataviensis fundatoris Gottwicensis et beati Gebehardi archiepiscopi Salisburgensis fundatoris monasterii Admontensis Benedictini in Styria ex **manuscriptis** vetustis **codicibus** editae neque levis momenti res suggerentes continentur.

BP 2 (Bernhard Pez an Moritz Müller; 1712-07-16 [?]; Wien; Edition 2010: Brief 259):

In quodam monasterio (videlicet Beatae Mariae Virginis ad Scotos Viennae ordinis nostri) reperi tres **libros** De viris illustribus monasterii S. Galli **manuscriptos**, quos uti suspicor esse Mezleri vestri, ita ut aurum habeo. Scriptus est ille **codex** anno 1606.

BP 3 (Bernhard Pez an René Massuet; 1712-08-21; Melk; Edition 2010: Brief 266):

Ubi enim illa diplomata, **chartae**, probationes, fides **manuscriptorum codicum** etc.?

BP 4 (Bernhard Pez an Sebastian Schott; 1714-01-07; Melk; Edition 2010: Brief 323):

Manuscriptos codices nullos inde accepi, nec habere admodum delectabat, utpote qui communes, passim vulgati, denique proletarii nec cum Tyrnstainensibus comparandi erant.

BP 5 (wie BP 4):

Manuscripti codices hic excepto uno unico, qui saeculo XV. scriptus erat, nulli.

BP 6 (Bernhard Pez an Moritz Müller; 1714-01-11; Melk; Edition 2010: Brief 325):

Nuntia, an **manuscriptos codices** habeant etc.

BP 7 (Bernhard Pez an René Massuet; 1714-09-06; Melk; Edition 2010: Brief 354):

Interea noveris me totum in excutiendis Germaniae bibliothecis, maxime **codicibus** earum **manuscriptis**, versari.

BP 8 & 9 (wie BP 7):

Una bibliotheca Mellicensis, in qua **codices manuscripti** prope MCC exstant, quae non suggerit? Ibi Praedicatio sancti Bonifacii Moguntini nondum edita in **codice** saeculi noni, si quid horum intelligo.

BP 10 (Bernhard Pez an Johann Christoph Bartenstein; 1714-09-20; o.O.; Edition 2010: Brief 360):

Cum enim uterque nostrum in Graecis versati simus, egoque evolvendis **codicibus manuscriptis** multo iam tempore me addixerim, quidnam id, quaeso, futurum sit, quod nostro Parisiensibus amicis gratificandi studio posset intercedere?

BP 11 (wie BP 10):

Eum tecum caesareae bibliothecae praefectus, vir totus ex humanitate factus, non aegre ad separatam bibliothecae cubiculum admittet, ubi singulis fere diebus duarum minime horarum spatio presbyter exscribendo **codici** vacare possit.

BP 12 (wie BP 10):

Interea hodie singularem ad dominum Gentillotium dabo epistolam, in qua nulla non machina adhibita eo virum percellere studebo, ut facultatem **codicis** e bibliotheca efferendi in tuamque domum asportandi indulgeat.

BP 13 (wie BP 10):

Nihil insolitum est, quod **libri** quidam et **codices** data syngrapha inde efferantur, id quod et ego paucis abhinc annis expertus fui.

BP 14 (wie BP 10):

Codex Graecus Nicephori Callisti olim in Galliam missus nec nisi aegerrime receptus multis nunc eruditorum votis negotium facessit.

BP 15 (Bernhard Pez an René Massuet; 1714-12-28; Melk; Edition 2010: Brief 379):

Nimirum totus ab eo tempore in bibliothecarum perlustratione fui tantumque non pulveribus **codicum**, quos usque ad valetudinis etiam periculum excussi, sepultus delitui.

BP 16 (wie BP 15):

Origenis Ἐξηγήσιν εἰς τοὺς Ψαλμοὺς ex pessimae notae **codice** caesareo, in quo me teste vix umbra literarum subinde deprehendi poterat, felicissime, quoad licuit, exscripsit.

BP 17 (wie BP 15):

At alii ego viae insisto. Excutio praeprimis **codices manuscriptos** indeque sexcenta nostrorum scriptorum nomina eruo, quae frustra in editis quaeras.

BP 18 (wie BP 15):

Acta haec ab Eynwico virginis confessario et praeposito conscripta ex coaevo nostrae bibliothecae **codice** eruo et primus edo, integramque Florianensis monasterii historiam praefigo.

BP 19 (wie BP 15):

Opus hoc viris eruditis eo nomine pergratum fore arbitror, quod praecipuos non unius Mellicii, sed et plerorumque Austriae monasteriorum **manuscriptos codices** inibi legent.

BP 20 (Bernhard Pez an Hieronymus Übelbacher; 1715-04-25; Melk;
Edition 2010: Brief 399):

Interim vel ex praesenti opusculo patet pulverulentos nostros in
excutiendis veteribus **codicibus** labores suo fructu et insigni
utilitate nequaquam carere.

BP 21 (wie BP 20):

Utinam clarissimus dominus Sebastianus mihi cum venia reverendissimae
ac perillustris dominationis vestrae mitteret duos **codices**
manuscriptos, quorum unus **in folio** vetus quoddam Chronicon Austriae
Germanice scriptum, alter vero **in quarto** Vitam beati Hartmanni primi
praepositi Claustroneoburgensis complectitur.

BP 22 (wie BP 20):

Commendavi hos **codices** singulariter domino Sebastiano,
dum Tyrnsteinii fui, et si recte memini, infixam desuper candidam
schedulam notavi, ut eos reperire adeo difficile illi non accideret.

BP 23 (Bernhard Pez an Anton Steyerer; 1715-12-12; Melk; Edition 2010:
Brief 470):

His diebus forte fortuna in quendam nostrae bibliothecae **codicem** incidi,
in cuius antico assere schema stirpis Austriaco-Habsburgicae manu
saeculi XV. depictum et exaratum erat.

BP 24 (Bernhard Pez an NN in St. Pantaleon zu Köln; 1709-12-26; Melk;
Edition 2010: Brief 23):

Ubi studuerit, quid (etsi parum) scripserit, et an typis vulgatum fuerit,
an vero inter **manuscripta** duntaxat delitescat.

BP 25 (Bernhard Pez an Ansgar Grass; 1710-12-17; Melk; Edition 2010:
Brief 134):

Reliquit **manuscriptum** adhuc in bibliotheca Corbeiensi superstes
De fide, spe et charitate, laudato antistiti suo Karoli Magni ex sorore
nepoti humiliter inscriptum, quod certe lampada olet praemorsosque
sapit unguis.

BP 26 (Bernhard Pez an René Massuet; 1711-07-11; Melk; Edition 2010:
Brief 177):

Ex monasterio Nideraltahensi nuper accepi memorabilia monasterii
Mettensis in Bavaria **manuscripta**, ope et diligentia reverendi domini
patris Placidi monachi Nideraltahensis, viri rerumstrarum studiosissimi.

BP 27 (wie BP 3):

Ibi Andreae Chesnii tomi, Labbé Bibliothecae **manuscriptorum** tomi
duo, dissertationes etc., Auberti Miraei Bibliotheca etc., ita ut homo
vix crederet, in hunc orbis eruditi angulum adeo nobiles et ex
remotissimis regnis allatos **libros** potuisse confluere.

BP 28 (Bernhard Pez an Moritz Müller; 1713-04-13; Melk; Edition 2010:
Brief 304):

Quaeso, nunciate, an et quae quotve **manuscripta** ibi conserventur?

BP 29 (Bernhard Pez an Moritz Müller; < 1714-04-07; o.O.; Edition 2010:
Brief 335):

De **manuscripto** Mezleri vestri, quod in Scotensi bibliotheca reperi,
nihil ausim pro certo affirmare, quin tibi ἀυτόγραφα omnia possidenti
unice assentiar.

BP 30 (Bernhard Pez an Moritz Müller; 1714-06-28; Melk; Edition 2010:
Brief 346):

Si quaedam de aliis historicis, maxime veteribus, qui subinde ex
manuscriptis eruti prodeunt, intellexeris, obsecro, nuntia.

BP 31 (wie BP 1):

In ea simul omnia, quae ad singula cuiusque in Suevia ordinis monasteria, casus, hodiernum statum etc. pertinent, barbara quidem oratione, sed vera, et quae multas egregias monasteriorum **chartas**, quas vocant, referat, descripta reperies.

BP 32 (wie BP 3):

Eo opusculo vir bonus celebrat laudes sui coenobii, sed ut dixi, ῥητορικῶς tantum, ignarus nos eruditosque omnes huiusmodi nugas pueris dudum cessisse oculosque duntaxat in firma illa et solida, qualia ex diplomatibus, **chartis** vetustioribus etc. eruuntur, defixos habere.

BP 33 : siehe BP 3.

BP 34 & 36 (wie BP 4):

Atlas minor, continens 30 circiter **chartas** seu mappas geographicas, **in folio**.

BP 35 (Bernhard Pez an René Massuet; < 1715-07-05; o.O.; Edition 2010:

Brief 411):

In monasterio Austriae Seittenstadiensi vidi Chronicon **manuscriptum** loci a saeculo XII. usque ad XVII. unacum litteris et **chartis** donationum etc.

BP 36: siehe BP 34.

BP 37 (wie BP 4):

Pro re bibliothearia [*sic*]: Guilielmi Cave Historia litteraria **in folio**, Genevae anno 1705.

BP 38 (wie BP 4):

Simleri Epitomen Gesneri, quam Tyrnstainium attuli quamque ipse vidisti, **in folio**.

BP 39 (wie BP 30):

Venerabilis Guiberti abbatis Novigentensis **folio** fl. 5.

BP 40: siehe BP 21.

BP 41 (wie BP 1):

Habeo penes me bina huius operis exemplaria **in quarto**, uti loqui mos est.

BP 42 (wie BP 3):

Est id typis editum Labaci anno 1689 **in quarto** authore Josepho N.
eiusdem coenobii priore.

BP 43 (wie BP 30):

Venerabilis Petri Cellensis **quarto** fl. 2.

BP 44: siehe BP 21.

BP 45 (wie BP 35):

Editus fuit hoc ipso anno Ratisbonae **in quarto**, quem cum iterum nactus
fuero (nam ad alienas primum exemplum manus iam avolavit), tuus erit.

BP 46 (wie BP 4):

Pro geographia: Mellisantis weldbeschreybung **in octavo**, duae partes.

BP 47 (wie BP 4):

Pro historia: Titus Livius. Philippi Brietii Annales mundi **in octavo**, duae
partes.

BP 48 (wie BP 4):

Ante omnia emes tibi que procurabis Die k rchenhistory Ludovici Du Pin
in octavo, duae partes.

BP 49 (Bernhard Pez an Adalbert Defuns; 1715-12-25; Melk; Edition 2010: Brief 476):

Caeterum maxime nunc opto, ut Epistolas meas apologeticas, quas adversus Jesuitam quemdam Viennensem in sacrum ordinem nostrum perquam contumeliosum hoc anno Campoduni **in octavo** edidi, tibi zelotissimo honoris Benedictini vindici offerre possim.

BP 50 (Bernhard Pez an NN in Břevnov; 1709-09-22; Melk; Edition 2010: Brief 7):

Constitui nuper Bibliothecam Benedictinam scribere, hoc est, tredecim fere **libris** scriptores Benedictinos omnes, quotquot ab ordine condito ad hanc usque aetatem nostram aliqua scribendi laude floruerunt, complecti.

BP 51 (wie BP 50):

[N]on dubito in eodem quoque asceterio quam plurimos extitisse viros praeclaros, qui editis etiam **libris** suum posteris nomen reliquerint.

BP 52 (Bernhard Pez an die Bayerische Benediktinerkongregation; 1709-10-23; Melk; Edition 2010: Brief 16):

Institueram confecto cursu theologico comparataque magna ex parte Graecae Hebraicaeque linguae scientia Bibliothecam Benedictinam scribere, id est: tredecim fere **libris** scriptores Benedictinos omnes, quotquot a sacro ordine nostro condito ad hanc usque aetatem ubicunque floruerunt, complecti.

BP 53 (Bernhard Pez an NN in Monte Cassino; 1709-11-17; Melk; Edition 2010: Brief 19):

Constitui pauco abhinc tempore Bibliothecam Benedictinam scribere, hoc est tredecim fere **libris** scriptores Benedictinos omnes, quotquot a sacro nostro ordine condito ad hanc usque aetatem ubicunque floruerunt, complecti.

BP 54 (wie BP 24):

Constitui sat multo abhinc tempore Bibliothecam Benedictinam scribere, hoc est, tredecim fere **libris** scriptores Benedictinos omnes, qui a primis sancti ordinis nostri incunabilis ad hanc usque aetatem aliqua scribendi laude claruerunt, complecti.

BP 55 (Bernhard Pez an die Maurinerkongregation; 1709-12-29; Melk; Edition 2010: Brief 24):

Constitui coepique multo abhinc tempore Bibliothecam Benedictinam scribere, hoc est, tredecim fere **libris** scriptores Benedictinos omnes, qui a sacro ordine nostro condito ad hanc usque aetatem ubicunque locorum floruerunt, complecti.

BP 56 (Bernhard Pez an Quirin Millon; 1710-01-07; Melk; Edition 2010: Brief 28):

[P]ossem in meam defensionem adducere beatum Paulum Warnefridi, cognomento Diaconum, professum Cassinensem, qui sex **libris** historiam belli Longobardici complexus est.

BP 57 (Bernhard Pez an René Massuet; 1710-05-18; Melk; Edition 2010: Brief 68):

Quo in genere omnium Germanorum infortunatissimi Austriaci sunt, utpote ad quos **libri** in Gallia editi vel pacis, nedum belli tempore raro admodum et summo pretio exportantur.

BP 58 (wie BP 57):

[P]raeterquam quod, qui patria et vernacula lingua suos **libros** quantumvis perfectos conscribunt, necesse habeant nunquam non metuere, ne in imperiti et illepidi interpretis manum incidant, id quod magnis persaepe viris contigit.

BP 59 (Bernhard Pez an Benedetto Bacchini; 1710-09-07; Melk; Edition 2010:
Brief 108):

Habeo ex Wionis **libro** secundo Ligni vitae capite 62 et sequentibus
multos egregios ex illo celeberrimo monasterio viros, sed admodum
ieiune et parce ab eodem Wione descriptos.

BP 60 (wie BP 1):

Erat is mihi probe ex scriptorum Ottoburanorum catalogo ab admodum
reverendo patre Alberto Krez misso cognitus, verum tot epistolarum
libros (namque duos duntaxat pater Albertus memorabat) scripsisse
penitus ignorabam.

BP 61 (wie BP 1):

De aliis **libris**, qui mihi usui esse possint, tu me pro mutuo inter nos
amore monebis.

BP 62 (Bernhard Pez an Moritz Müller; 1712-02-13; Melk; Edition 2010:
Brief 209):

Porro, dum praeter egregium hunc erga me animum doctissima illa
librorum missorum munera quoque considero, omnino, quid
cogitem reponamve, ignoro.

BP 63 (wie BP 2):

Ego, ut hucusque saepius innui, nunc ad breve tempus Viennae ago,
scrutaturus bibliothecas fere omnes et maxime caesaream, in qua
optimorum **librorum** immensa multitudo.

BP 64: siehe BP 2.

BP 65 (wie BP 3):

[H]omo vix crederet, in hunc orbis eruditi angulum adeo nobiles et
ex remotissimis regnis allatos **libros** potuisse confluere.

BP 66 (wie BP 28):

His catalogis incredibile meum opus augetur. Mezlerum vestrum, id est tres **libros** De viris illustribus S. Galli manuscriptos ad manum habeo.

BP 67 (Bernhard Pez an Moritz Müller; 1713-07-16; Melk; Edition 2010: Brief 318):

Sarcinam interim omnem **librorum** Ulma, non sine summa omnium nostrum admiratione, tripudio, et tui erga nos studii depraedicatione, 10. huius accepimus.

BP 68 & 69 (wie BP 4):

In canonia Ducumburgensi quinque fere dies exegi, nec paenitendos inde **libros** Mellicensibus destinatos excerpti inque nostrum monasterium, remissis illuc aliis paris aut etiam maioris pretii **libris**, retuli.

BP 70 (wie BP 4):

Reliqui **libri** pro necessariis viro erudito disciplinis, quos tibi dictare nequivi, hi sunt [...].

BP 71 (wie BP 29):

Interim vide totius operis veluti oeconomiam. Titulus: De viris illustribus monasterii S. Galli **libri** tres. Anno Domini 1606.

BP 72 (wie BP 30):

De **libris** Gallicis submolestiora nuntias, licet verissima.

BP 73 (wie BP 30):

Tanta nunc ubique indignitas iniquitasque hominum rerumque est, ut vel **librorum** meminisse non raro terrori sit.

BP 74: siehe BP 13.

BP 75 (wie BP 35):

De aliis, uti Othlonis **libris** tribus De providentia, nunc nihil adscribo, ne dolorem tuum, cum singula mittere aut exscribere nequiverimus, augeam.

BP 76 (wie BP 49):

Sed ego ita hominem accepi, ut omnes nunc Socii sentiant sodalis imprudentiam, quidam etiam in me prope furunt et **librum** meum, quocunque modo possunt, supprimunt [...].

BP 77 (wie BP 56):

Denique me merito absolvat Abbo Parisiensis in monasterio ad S. Germanum a Pratis prope Parisios professus, qui saeculo Christi nono obsidionem Parisiensem duobus **libellis** carmine heroico contextis cecinit.

BP 78: siehe BP 1.

BP 79 (wie BP 3):

Egregium illum criticum Erathium mirum est, quam belle et festive in tuo **libello** seu epistola habueris.

BP 80 (wie BP 28):

Porro mei **libelluli** tanti sunt haudquaquam, ut magno admodum eorum desiderio teneamini.

BP 81 (wie BP 29):

In hac praefatione ad calcem fere graviter Mezlerus conqueritur, quod quidam viri docti opusculum hoc (quod se adversa valetudine detentum tumultuarie confecisse praemiserat) perviderint, extorserint, imo rapuerint, et se inscio Ingolstadianis typis **libellum** tertium totum, multa ex primo, et ex secundo de [†].

BP 82 & 83 (wie BP 35):

Ausus fuit nimirum ex Societate anonymus quidam contumeliosum nostro ordini **libellum** Viennae proximo anno in ipsa universitatis Viennensis luce vulgare. Igitur cum nostrum silentium ipso **libello** propudiosius fore mei superiores arbitrati sint, mihi in mandatis dedere, ut ita homini profecto ignoranti responderem [...].

BP 84 (wie BP 49):

Ansam huic epistolae praebuit haec mendax Jesuitae propositio **libello** publice edito comprehensa [...].

BP 85 (Bernhard Pez an Othmar Zinke; 1710-10-04; Melk; Edition 2010: Brief 113):

[...] eruditissimi nostri congregationis sancti Mauri in Gallia patres, qui, cum Annales Benedictinos a celeberrimo illo Galliae phoenice Joanne Mabillione coeptos et usque ad annum Domini 1011 quatuor bene magnis **voluminibus** explicatos ad optatum denique exitum perducendos suscepissent [...].

BP 86 (wie BP 26):

Iam conclusi epistolam, cum ecce abbas S. Crucis ex Polonia in meo monasterio adest. Is omnia fere, quae petis, sat grandibus **manuscriptis voluminibus** comprehensa habebat.

BP 87 (wie BP 30):

Bibliothecam patrum Lugdunensem quominus meo nomine emas, obstant sex vasta **volumina** Scriptorum Francicorum et Nortmannicorum Du Chesne, fl. 150 nuper a me comparata.

BP 88 (wie BP 30):

Sancti Ambrosii **volumina** duo compacta fl. 42.

BP 89 (wie BP 15):

Iam vero quod scribis quendam Hispanum eandem quam ego spartam
adornare et nescio quae iam Bibliothecae Benedictinae **volumina**
edidisse, me non admodum commovet.

7) Anhang 3: Abstract

Vorliegende Arbeit versammelt Erkenntnisse, die im Zuge computergestützter linguistischer Annotation ausgewählter frühneuzeitlicher Gelehrtenbriefe neulateinischer Sprache gewonnen werden konnten. Bei diesen Texten handelt es sich unter anderem um die Briefe des ursprünglich protestantischen, später katholischen Historikers und Bibliothekars Johann Georg Eckhart (gest. 1730) an die Melker Geschichtsforscher Bernhard und Hieronymus Pez OSB (gest. 1735 bzw. 1762).

Die Absicht vorliegender Arbeit besteht dabei darin, das Potential digitaler, linguistischer Annotation für die Forschung anhand von neulateinischen Texten aufzuzeigen.

Digital-linguistische Annotation (linguistisches Tagging) stellt innerhalb der Corpuslinguistik sowie innerhalb der Digital Humanities ein anerkanntes Verfahren dar, das zumeist auf modernsprachliche Textsammlungen angewandt wird und das unter anderem auf die (vergleichende) sprachwissenschaftliche Auswertung (mehrerer) mitunter sehr großer Textquanten abzielt. Durch linguistische Annotation werden die in einem Text enthaltenen sprachlichen Informationen, z.B. Wortarten und morphologische Phänomene, explizit gemacht, maschinell les- und verarbeitbar. Innerhalb der Editions- und der Geschichtswissenschaft eröffnet digital-linguistische Annotation neue Horizonte für die Auswertung historischer Quellen: Es wird dadurch eine Ausgangsbasis geschaffen für diachrone, sprachliche Vergleiche älterer Sprachstufen, was, verglichen mit modernsprachlichen Corpora, im deutschsprachigen Raum bisher seltener der Fall war. Ferner kommt linguistische Annotation der systematischen Durchsuchbarkeit schriftlicher Quellen in höchstem Maße zugute. Ihrem Prinzip nach kann linguistische Annotation sprach- und quellenunabhängig vorgenommen werden.

Vorliegende Arbeit befasst sich mit dem Begriff, dem aktuellen Forschungsstand, der Praxis und den gängigsten Varianten linguistischer Annotation und wendet ausgewählte Teile ihres Methodenspektrums speziell auf neulateinische Gelehrtenbriefe in digitaler Form an. Im Zentrum steht dabei die Frage, welche Standards und welche Varianten linguistischer Annotation in Hinblick auf die digitale Edition neulateinischer Quellen der frühen Neuzeit am besten geeignet

erscheinen. Zur Diskussion stehen einerseits linguistische Basisannotation mit Kennzeichnung der Wortarten und mit Ausweis der Grundworte (Lemmatisierung), andererseits morphologische Annotation und syntaktisches Treebanking. Ferner werden Annotationsschemata (Tagsets) evaluiert, die im Rahmen einschlägiger Projekte der jüngeren Vergangenheit auf Texte lateinischer Sprache angewandt wurden (v.a. im Rahmen der Perseus Digital Library sowie des Index Thomisticus Online). Neben der Bereitstellung digitaler Infrastruktur bieten diese Projekte wenig Auswertungen des eigenen Materials, durch die das Potential linguistischer Annotation verdeutlicht würde.

Auf Basis der annotierten Briefe bietet vorliegende Arbeit semantische Analysen ausgewählter Begriffe sowie von deren unmittelbarem, sprachlichem Kontext: Analysiert wird unter anderem die Verwendung von Termini wie *codex*, *liber* oder *manuscriptum*. Diese Begriffe sind innerhalb der Korrespondenz zwischen Johann Georg Eckhart, Bernhard Pez und anderen Korrespondenzpartnern von zentraler Bedeutung für deren Interessen als Geschichtsforscher.

Die im Zuge der praktischen Annotation generierten Daten, die annotierten Briefe in digitaler Form, sind am Server des Austrian Centre for Digital Humanities der Österreichischen Akademie der Wissenschaften in Wien gespeichert. Interessierten Forscherinnen und Forschern können sie auf Anfrage auch zur weiteren Bearbeitung und Auswertung zugänglich gemacht werden.

Jenes webbasierte Tool, mit dem die Annotation durchgeführt wurde, wird in vorliegender Arbeit ausführlich vorgestellt und beschrieben – der vom ACDH der ÖAW in Wien entwickelte TokenEditor.