Ein digitales Textformat für die Literaturwissenschaften

Die Richtlinien der Text Encoding Initiative und ihr Nutzen für Textedition und Textanalyse

Christof Schöch (Würzburg)

zusammenfassung: Die stetig voranschreitende Digitalisierung literarischer Texte verschiedenster Sprachen, Epochen und Gattungen stellt die Literaturwissenschaften immer wieder vor die Frage, wie sie diese Entwicklung mitgestalten und zu ihrem Vorteil nutzen können. Dabei ist digital nicht gleich digital, sondern es existiert eine Vielzahl sehr unterschiedlicher, digitaler Repräsentationsformen von Text. Nur wenige dieser Repräsentationsformen werden literaturwissenschaftlichen Anforderungen tatsächlich gerecht, darunter diejenige, die den Richtlinien der Text Encoding Initiative folgt. Der vorliegende Beitrag vergleicht zunächst einige derzeit gängige digitale Repräsentationsformen von Text. Für literaturwissenschaftliche Forschung besonders geeignet erweist sich hierbei eine Repräsentationsform, die den Richtlinien der Text Encoding Initiative folgt. Daher informiert der Beitrag anschließend über deren Nutzen für die literaturwissenschaftliche Arbeit, sowohl im Bereich der wissenschaftlichen Textedition als auch im Bereich der Analyse und Interpretation von Texten. Nur wenn die Literaturwissenschaften in ihrer Breite den Nutzen von offenen, expressiven, flexiblen und standardisierten, langfristig nutzbaren Formaten für die Forschung erkennen, können sie sich mit dem erforderlichen Nachdruck für deren Verbreitung einsetzen und durch die zunehmende Verfügbarkeit von Texten in solchen Formaten für die eigene Forschung und Lehre davon profitieren.

schlagwörter: Digital Humanities; Text Encoding Initiative; Textedition; Textanalyse

Einleitung

Eine der zentralen Einsichten der Literaturwissenschaften im 20. Jahrhundert war die analytische Unterscheidung von Form und Inhalt bei gleichzeitigem Bewusstsein ihrer untrennbaren Verbundenheit und gegenseitigen Abhängigkeit.[1] Das medientheoretische Analogon dieser Einsicht formulier-

[1] Ein Beispiel für diese Position ist schon Jean Rousset, *Forme et signification: essais sur les structures littéraires de Corneille à Claudel* (Paris: Corti, 1962).

te Marshall McLuhan in seinem Diktum „The medium is the message".[2] Und die Editionswissenschaften haben sich ausführlich der Frage gewidmet, wie sich Manuskripte, Typoskripte, Druckfahnen und verschiedene Textausgaben zueinander verhalten und damit auch die Frage nach der Beziehung zwischen den materiellen Trägern des Textes und seiner Überlieferungsgeschichte untersucht.[3] So ist in unterschiedlichen Bereichen immer wieder deutlich geworden, wie eng Inhalt, Form und Medium zusammenhängen. Mit dem sich seit den 1960er Jahren entwickelnden, seit den 1990er Jahre rasant an Fahrt gewinnenden digitalen Paradigmenwechsel in Gesellschaft und Wissenschaft ist ein neuer Aspekt der medialen Realisierungsformen (literarischer) Texte hinzugekommen, der erst in jüngerer Zeit in das Blickfeld des Interesses gerückt ist: Welchen Unterschied macht es, wenn (literarische) Texte nicht in Form von Handschriften oder gedruckten Büchern, sondern (wie dies zunehmend der Fall ist) in Form von Dateien, also digitalen Textdaten, vorliegen? Wie kann ein ursprünglich gedruckt erschienener Text adäquat ins digitale Medium überführt werden, und welche digitalen Repräsentationsformen sind verfügbar? Inwiefern spielt die jeweilige digitale Repräsentationsform eine Rolle für die Rezeption und Interpretation eines (literarischen) Textes? Welche Möglichkeiten und Herausforderungen eröffnet das Vorliegen digitaler Texte für die Bearbeitung literaturwissenschaftlicher Fragestellungen, und wie verändern sich die hierfür eingesetzten Methoden?

Der vorliegende Beitrag möchte einige Aspekte dieser Fragen diskutieren. Er wird hierfür zunächst darauf fokussieren, welche unterschiedlichen digitalen Repräsentationsformen literarischer Texte den Literaturwissenschaften derzeit zur Verfügung stehen. Mit dem Begriff ‚Repräsentationsform' sind hier unterschiedliche Datenformate gemeint, die der Speicherung, Darstellung und Analyse von Texten dienen. In dieser Perspektive werden zunächst einige unterschiedliche digitale Repräsentationsformen von Text beschrieben und miteinander verglichen (Abschnitt 1.1). Anschließend argumentiert der Beitrag, dass eine adäquate, digitale Repräsentationsform von literarischen Texten eine ganze Reihe von Anforderungen erfüllen sollte, um nicht nur dem Gegenstand selbst, sondern auch verschiedenen literaturwissenschaftlichen Nutzungsszenarien mit ihren Erkenntnisinteressen, Frage-

[2] Marshall McLuhan, *Understanding Media: The Extensions of Man (1964)*, hrsg. von W. Terrence Gordon, Critical edition (Corte Madera: Gingko Press, 2003).

[3] Unter anderen: Almuth Grésillon, *Eléments de critique génétique: lire les manuscrits modernes* (Paris: Presses universitaires de France, 1994).

stellungen und Methoden gerecht zu werden. Es wird deutlich, dass die derzeit am Besten geeignete Repräsentationsform die von der Text Encoding Initiative (TEI) vorgeschlagene ist. Inwieweit ein nach den Richtlinien der TEI kodierter Text die zuvor beschriebenen Anforderungen erfüllt, wird daher ebenfalls gezeigt (Abschnitt 1.2). Bevor die Nutzung von TEI in der literaturwissenschaftlichen Arbeit ausführlicher diskutiert wird, werden die Geschichte, Ziele und grundlegenden Eigenschaften von TEI-kodierten Dokumenten ebenfalls in aller Kürze dargestellt (Abschnitt 1.3). Wesentlich ist dabei, dass TEI technisch gesehen auf dem weit verbreiteten Standard XML (Extensible Markup Language) basiert; dass in TEI der Text selbst, Annotationen einzelner Passagen und dokumentbezogene Metadaten gemeinsam vorliegen; und dass die Richtlinien der TEI seit 1987 von einer wissenschaftlichen Gemeinschaft entwickelt werden.

Eine immer größer werdende Zahl wissenschaftlicher Digitalisierungs- und Editionsprojekte bietet ihre Texte unter anderem in Form von mehr oder weniger aufwändig nach den Richtlinien der TEI erstellten Dateien an. Warum es sich lohnt, den damit verbundenen, nicht immer unerheblichen Aufwand zu treiben, wird für zwei literaturwissenschaftliche Handlungsfelder anhand konkreter Beispiele aufgezeigt: Erstens für das Gebiet der digitalen Textedition, mithin für den Bereich der Textkonstitution selbst (Abschnitt 2). Und zweitens für das Gebiet der Textanalyse und Interpretation (Abschnitt 3). Für beide Bereiche wird auf Beispiele aus der deutschen, französischen und spanischen Literaturwissenschaft zurückgegriffen. Der Beitrag möchte hierbei aufzeigen, wie die Nutzung eines Standards wie TEI den Literaturwissenschaften einen dem Gegenstand und den literaturwissenschaftlichen Zielen angemessenen, differenzierten Zugriff auf die Texte ermöglicht.

Die Darstellung hat dabei einerseits einführenden Charakter und verfolgt das Ziel, grundlegend über die Texte Encoding Initiative und ihren Nutzen für die literaturwissenschaftliche Arbeit zu informieren. (In diesem Sinne bietet der Beitrag im Anhang auch Hinweise für die weitere Beschäftigung mit dem Thema an.) Zugleich versteht sich der Beitrag als Plädoyer für die Nutzung digitaler Repräsentationsformen von Text und insbesondere der Text Encoding Initiative als grundlegendes literaturwissenschaftliches Arbeitsinstrument. Das Thema, so die hier vorgetragene Überzeugung, betrifft nicht nur diejenigen Forschenden, die aktuell ihrem Selbstverständnis nach als Digitale EditionswissenschaftlerInnen oder ComputerphilologIn-

nen agieren, sondern zunehmend große Teile der Literaturwissenschaften insgesamt. Der Beitrag möchte in diesem Sinne nicht nur aufzeigen, wie eng die Beziehungen zwischen Textkonstitution und Textanalyse sein können, sondern vor allem auch einen Beitrag zum Brückenschlag zwischen digitalen Geisteswissenschaften und etablierten Literatur- und Kulturwissenschaften leisten.

1. Digitale Repräsentationsformen literarischer Texte

Die Digitalisierung des kulturellen Erbes wurde in den letzten Jahrzehnten von vielfältigen Akteuren vorangetrieben, darunter zahlreiche Bibliotheken (man denke als herausragendes Beispiel an Gallica, die Digitalisierungsinitiative und Plattform der Bibliothèque nationale de France), Forschungsprojekte (wie unter vielen anderen TextGrid oder die Deutsche Digitale Bibliothek) sowie privatwirtschaftliche Akteure (hier drängt sich das nicht unumstrittene Digitalisierungsprogramm Google Books als Beispiel auf). Diese Aktivitäten haben eine Vielzahl digitaler Repräsentationsformen von (literarischen) Texten hervorgebracht unter denen diejenigen, die bei der literaturwissenschaftlichen Arbeit am häufigsten begegnen, hier zunächst charakterisiert werden sollen, bevor sie mit literaturwissenschaftlichen Anforderungen konfrontiert werden.

1.1 Einige verbreitete, digitale Repräsentationsformen literarischer Texte

Für die Einordnung dieser Repräsentationsformen und der entsprechenden Dateiformate können die folgenden drei grundlegende Beschreibungsdimensionen genutzt werden. Erstens: Ist das Format proprietär oder offen? Das heißt, ist frei nutzbar, einsehbar und modifizierbar oder ist das nicht der Fall, beispielsweise weil das Format von einer Firma kontrolliert wird? Zweitens: Liegen die Informationen in dem Format wenig strukturiert, semi-strukturiert oder relativ stark strukturiert vor? Und drittens: Ist das Format auf die Darstellung der Texte hin orientiert, oder steht die Repräsentation von Struktur und Semantik der Texte im Vordergrund?

1. PDF: darstellungsorientiertes, offenes, wenig strukturiertes Format. PDF-Dateien sind ein Containerformat, das sowohl Bild- als auch Textinformationen beinhalten kann. Literarische Texte (und vieles mehr, unter anderem wissenschaftliche Aufsätze) begegnen sehr häufig im PDF-Format, weil dieses nach wie vor das übliche Distributionsformat der meisten Bibliotheken ist. Durch das Scannen einer handgeschriebenen oder gedruckten Vorlage enstehen

Bilddateien, die meist als hochauflösende TIFF-Datei archiviert und den NutzerInnen in Form von PDF-Dateien zum Download angeboten werden. Nur wenn OCR (Optical Character Recognition) angewandt wird, kann auch der im Bild sichtbare Text in die PDF-Datei eingebettet und dann durchsucht werden. Die Dateien benötigen im Vergleich zu reinen Textformaten relativ viel Speicherplatz, was bei größeren Textmengen ein Nachteil sein kann. Ein Vorzug der beschriebenen PDF-Dateien liegt in der originalgetreuen und unveränderlichen Wiedergabe von Layout und anderen visuellen Eigenschaften von Dokumenten. In der Regel erheben die Bibliotheken auch Metadaten, also Informationen, die das Dokument beschreiben und kontextualisieren.

2. TXT: offener, unstrukturierter Volltext. Dieser sogenannte *plain text*, der nur minimalen Informationen über die Zeichenfolge hinaus enthält, ist eine der am weitesten verbreiteten Formen digitaler (literarischer) Texte. (Nicht nur das *Gutenberg Project*, auch das *Internet Archive* und viele Bibliotheken, darunter *Gallica*, die digitale Bibliothek der französischen Nationalbibliothek, bieten solche Textdateien an.) Solche Volltexte sind in der Regel das Ergebnis einer OCR-Prozedur, die den Textinhalt aus einem digitalen Faksimile extrahiert, oder entstehen durch manuelle Transkription. Dieses Format hat gegenüber dem reinen digitalen Faksimile als Bilddatei den Vorteil, dass der Text als Zeichenfolge vorliegt und durchsucht werden kann und nur wenig Speicherplatz benötigt. Es finden sich allerdings Texte in sehr unterschiedlicher Güte, die vom Digitalisierungsmodus abhängen.[4] Ein Vorteil von *plain text* ist, dass die Dateien mit diversen Programmen bearbeitet werden können. In Abhängigkeit von der gewählten Zeichenkodierung können Zeichen, die über den Grundbestand des englischen Alphabets hinausgehen (wie sie für viele europäische Sprachen wichtig sind), nicht (bei reinem ASCII), nur auf einer bestimmten Plattform (ANSI) oder aber sehr vollständig (bei UNICODE/UTF-8) repräsentiert werden.[5]

[4] Die OCR-Bearbeitung großer Textmengen ohne manuelle Nachbearbeitung ist vor allem für Dokumente, die vor 1800 publiziert wurden, immer noch äußerst fehlerbehaftet. Nützliche Hinweise zur Textdigitalisierung bietet: DFG, „DFG-Praxisregeln Digitalisierung" (Bonn: Deutsche Forschungsgemeinschaft, 2013), www.dfg.de/formulare/12_151.

[5] Zu den genannten Kodierungsformaten und der Problematik der Zeichenkodierung allgemein, siehe Joel Spolsky, „The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!)", *Joel on Software*, 2003, www.joelonsoftware.com/articles/Unicode.html.

3. DOCX, ODT und HTML: darstellungsorientierter, wenig strukturierter Volltext. Diese Textformate sind allgemein sehr weit verbreitet, weil sie das übliche Format für Texte ist, die mit Office-Programmen wie Microsoft Word (Dateiformat DOCX, proprietär) oder LibreOffice (ODT, offen) erstellt werden bzw. weil sie im World Wide Web ubiquitär sind (HTML, offen), woraus sich auch eine gewisse Vertrautheit mit den Formaten ergibt. Gegenüber dem einfachen Volltext kann hier etwas mehr Strukturinformation festgehalten werden, vor allem wenn hierfür nicht eine rein visuelle Darstellung (Fett- oder Kursivschreibung, Textgröße, Schriftart), sondern Formatvorlagen systematisch eingesetzt werden. Dies ist allerdings häufig nicht der Fall, denn es handelt sich (auch bei HTML) um Formate, die auf die Darstellung hin entwickelt worden sind. Auf diese Weise erstellte Texte haben eine begrenzte Expressivität, was Informationen über die Struktur des Textes und die Eigenschaften von Einzelwörtern betrifft. Zudem wird ein Text, der mit Microsoft Word auf einem Windows-basierten Computer erstellt wurde, nicht unbedingt korrekt dargestellt, wenn er mit LibreOffice auf einem Linux-basierten Computer geöffnet wird. Dieses Kompatibilitätsproblem verschärft sich noch, wenn man bedenkt, dass eine heute sorgfältig erstellte Datei ja auch in zehn oder zwanzig Jahren noch verwendbar sein soll. Gut nutzbare Metadaten sind bei solchen Formaten in der Regel nur rudimentär vorhanden.

4. XHTML und EPUB: offener, darstellungsorientierter, semi-strukturierter Volltext. Eine weitere, stark verbreitete Form digitaler Texte sind nach offenen Standards erstellte, semi-strukturierte Textdokumente. Hierzu gehören alle Formate, die auf XML (Extensible Markup Language) basieren, so insbesondere XHTML (die etwas strengeren Regeln folgende Variante des weit verbreiteten, sehr einfachen HTML-Formats) und das offene Ebook-Format EPUB (das intern ebenfalls auf XHTML beruht und von der Mehrheit der Ebook-Anbieter und Lesegeräte unterstützt wird). Die Möglichkeiten, detailliert und systematisch Strukturinformationen über die Texte festzuhalten sowie lokale Annotationen und dokumentbezogene Metadaten einzubinden, sind relativ begrenzt. Da es sich um reine Textdateien nach den Prinzipien von XML handelt, können die Daten auch in Zukunft plattformunabhängig verwendet werden, zudem bleiben die Dateigrößen relativ gering.

5. TEI: offener, nicht-darstellungsorientierter, semi-strukturierter Volltext. Wie XHTML beruht auch das TEI-Format auf XML. TEI unterscheidet sich von XHTML aber entscheidend in zwei Punkten. Erstens ist TEI fast vollständig auf die Erfassung der Struktur von Texten und die abstrakten Eigenschaften

von Textabschnitten hin orientiert, das die Frage der Darstellung hiervon vollständig abkoppelt. (Für die Darstellung werden TEI-Dateien automatisch in entsprechende darstellungsorientierte Formate transformiert, darunter HTML, EPUB oder PDF.) Und zweitens ist es ein wesentlich expressiveres Format als XHTML, das heißt eine sehr große Anzahl relevanter Textphänomene kann in diesem Format explizit gemacht werden und detaillierte Metadaten können festgehalten werden. Auch hier gilt, dass die technische Grundlage XML ist und die Daten plattformunabhängig auch langfristig verwendet werden können. TEI ist ein offenes Format, das von einer Institution, der Text Encoding Initiative, getragen und gepflegt wird.

6. XMI und TCF: offener, nicht-darstellungsorientierter, stark strukturierter Volltext. Eine spezielle, noch stärker strukturierte Form der Textrepräsentation, die allerdings enge Bezüge zu der zuletzt genannten Form aufweist, sind Datenformate aus der Computerlinguistik, die insbesondere für die detaillierte, linguistische Annotation von Texten auf mehreren Ebenen entwickelt wurden und seit einiger Zeit neben einfachere, tabellarische Formate getreten sind. Es gibt zahlreiche entsprechende Formate, von denen hier nur zwei charakterisiert werden sollen: Erstens das sehr generische XMI (XML Metadata Interchange), das den eigentlichen Text von den linguistischen Annotationen trennt und beide über ein System von Identifikatoren verbindet. Dies ist ein sehr mächtiger und flexibler Mechanismus, der insbesondere den Austausch von Annotationen aus verschiedenen Quellen ermöglichen soll, der aber auch zu einer gewissen Komplexität beim Prozessieren der Informationen führt.[6] Und zweitens das von CLARIN-D entwickelte TCF (Text Corpus Format), das auf XML basiert und für jede Annotationschicht (Lemmata, Wortarten, syntaktische Dependenz, Named Entities, etc.) einen eigenen Dateibereich vorsieht und die sprachlichen Einheiten ebenfalls über Identifikatoren verknüpft.[7] Beide Formate sind offene, sehr expressive Standards, die sich auch für die langfristige Speicherung von Daten eignen.

Es stellt sich nun die Frage, welches dieser Formate für die Repräsentation von (literarischen) Texten am besten geeignet ist, wenn in erster Linie literaturwissenschaftliche Anforderungen und Nutzungsszenarien berücksichtigt werden sollen, und welche Anforderungen dies sind.

[6] Zu XMI, siehe Kapitel 4 in: Graham Wilcock, *Introduction to Linguistic Annotation and Text Analytics*, 3 (San Rafael, Calif.: Morgan & Claypool, 2009).

[7] Zu TCF, siehe http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.

1.2 Anforderungen an eine literaturwissenschaftlich adäquate digitale Repräsentationsform literarischer Texte

In diesem Abschnitt werden einige der Anforderungen, welche literaturwissenschaftliche Nutzungsszenarien an eine digitale Repräsentationsform von literarischen Texten stellen, begründet und erläutert.[8]

1. Explizite und korrekte Repräsentation der Zeichensequenz. Die einzelnen Zeichen, aus denen der Text besteht, werden adäquat kodiert. Texte bestehen nicht nur aus den 26 Buchstaben des englischen Alphabets, sondern alle Zeichen aller (auch historischer) Sprachen müssen korrekt und präzise repräsentiert werden können. Diese Anforderung bedeutet, dass die Repräsentation in einer UNICODE-basierten Form kodiert werden sollten, beispielsweise in UTF-8. Nur so können die Texte dann auch automatisch nach korrekten Zeichensequenzen, bspw. Wörtern, durchsucht werden.

2. Repräsentation der Textstruktur. Nicht nur die Zeichen- und Wortsequenz, sondern auch die (hierarchische und anderweitige) Struktur des Textes sowie Aspekte des Layouts werden mit repräsentiert. Literarische Texte und die Gattungen, in die sie sich gliedern lassen, zeichnen sich unter anderem durch eine mehr oder weniger präzise formale Struktur (Sätze, Absätze, Kapitel und Teile bzw. Akte und Szenen oder Strophen, etc.) aus, die für Analyse und Interpretation wichtig sein kann. Auch layout-bezogene und typographische Eigenheiten wie Zeilen- und Seitenumbrüche können relevant sein. Auch Beziehungen zwischen mehr oder weniger großen Struktureinheiten und zwischen verschiedenen Texten sollten repräsentierbar sein.

3. Repräsentation editionswissenschaftlicher Phänomene. Editionswissenschaftlich konzipierte Textausgaben haben in der Regel den Anspruch, den Text nicht nur als strukturierte Zeichensequenz zu transkribieren, sondern auch

[8] Schon 1991 formuliert Michael Sperberg-McQueen einen Anforderungskatalog, der für die Entwicklung von TEI im Übrigen richtungsweisend war (Michael Sperberg-McQueen, „Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts“, *Literary and Linguistic Computing* 6, Nr. 1 (1991): 34–46, http://doi.org/10.1093/llc/6.1.34). Ein frühes literaturwissenschaftliches Plädoyer für die TEI ist Fotis Jannidis, „Wider das Altern digitaler Texte: philologische Textauszeichnung mit TEI“, *Editio*, Nr. 11 (1997): 152–77. Eine ausführliche Reflexion über Anforderungen an digitale Texte aus Sicht der Editionswissenschaften ist Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts* (Cambridge: Cambridge Univ. Press, 2006), www.cambridge.org/us/academic/subjects/literature/printing-and-publishing-history/gutenberg-google-electronic-representations-literary-texts. Hier soll darüber hinaus auch die Textanalyse explizit berücksichtigt werden.

editorisch relevante Phänomene explizit zu machen und eventuelle editorische Eingriffe, die zum Prozess der Textkonstitution gehören, zu dokumentieren. Darüber hinaus ist es häufig erforderlich, den in einem Manuskript erkennbaren Schreibprozess oder komplexe Beziehungen zwischen mehreren Textfassungen festzuhalten. Eine hierfür geeignete Repräsentationsform muss die theoretischen Annahmen verschiedener editorischer Schulen flexibel unterstützen.

4. Repräsentation linguistischer Information. Nicht nur die Einzelwörter, sondern auch ihre grundlegenden Eigenschaften (Grundform, Wortart, Wortfeld) werden mit repräsentiert. Literaturwissenschaftliche Analysen haben ein intensives Interesse nicht nur an den Einzelwörtern als Zeichensequenz, sondern vor allem daran, wie ihre Eigenschaften auf verschiedenen Ebenen zusammenspielen, um auf komplexe Weise Bedeutungskonstitution zu erlauben. Daher sollte es möglich sein, auch Informationen wie die Grundform, die Wortart, die syntaktische Funktion oder das Wortfeld, die durch linguistische Annotation gewonnen werden können, mit zu repräsentieren.

5. Repräsentation von Kontext. Nicht nur der Text selbst, sondern auch sein Kontext und seine Überlieferungsgeschichte (einschließlich der Entstehungsgeschichte des digitalen Dokuments selbst) werden mit repräsentiert. Die Bedeutung literarischer Texte ist in großem Maße auch von ihrem Entstehungskontext (Autor, Datum, Ort, Verlag, Epochenzugehörigkeit, Gattungszugehörigkeit, etc.) abhängig. Zudem ist wesentlich, welche Ausgabe verwendet wurde und welche Prinzipien der Textkonstitution dieser Ausgabe sowie der neuen, digitalen Version zugrunde lagen. Eine Repräsentationsform sollte also auch diese Informationen beinhalten, ohne dass sie aber mit dem eigentlichen Text vermischt werden.

6. Für Menschen und Maschinen lesbar. Der digitale Text ist für die individuelle, interpretierende Lektüre ebenso geeignet wie für die algorithmische Verarbeitung und Analyse. In Zukunft wird literaturwissenschaftliche Arbeit mit immer größerer Selbstverständlichkeit zwischen der intensiven Lektüre eines Einzeltextes, der einfachen Suche nach Stichworten, der händischen, qualitativen Annotation von Texten und der Exploration großer Textsammlungen mit quantitativen Verfahren wechseln.[9] Die Repräsentationsform sollte es daher mit möglichst geringem Aufwand ermöglichen,

[9] Also zwischen *close reading* und *distant reading* im Sinne Franco Morettis oder der Macroanalysis im Sinne Matthew Jockers'.

verschiedene Perspektiven auf den Text bzw. andere Repräsentationformen zu generieren, die sich für verschiedene Herangehensweisen und Nutzungsszenarien eignen. Hieraus ergibt sich die Notwendigkeit für eine flexible, informationsreiche, digitale Repräsentationsform, denn aus ihr lassen sich verschiedene, spezifischere Formate ohne größeren Aufwand generieren, während der umgekehrte Weg wesentlich aufwändiger ist.

7. Langfristige Nutzbarkeit. Der Text ist nicht nur heute, sondern auch in Zukunft nutzbar. Der nicht unerhebliche Aufwand, der in die Erstellung adäquater Textrepräsentationen investiert wird, sollte also nur einmal und in Zukunft nicht erneut notwendig sein, zumindest nicht von Grund auf. Hieraus ergibt sich die Notwendigkeit, sich in möglichst vielen Aspekte einer Repräsentationsform auf etablierte Standards zu berufen und keine Formate zu wählen, die an eine bestimmte Plattform, ein bestimmtes Betriebssystem, bestimmte Programme oder Firmen gebunden sind. Dies ist mit dem Begriff der technischen Interoperabilität gemeint, welche die Grundlage für die langfristige Nutzbarkeit eines Textformats darstellt. Für die langfristige Nutzbarkeit eines solchen offenen Standards ist wesentlich, dass nicht nur eine technische Spezifikation vorhanden ist, sondern der Standard auch von einer breiten Community getragen und von einer verantwortlichen Institution gepflegt wird.

Tabelle 1 zeigt, in welchem Maße die im vorigen Abschnitt beschriebenen Repräsentationsformen von Text (hier nach den erwähnten Dateiformaten unterteilt) die genannten Anforderungen erfüllen.

Kriterium	PDF	TXT	DOCX	XHTML	TEI	XMI	TCF
1. Repräsentation der Zeichensequenz	o	+	+	+	+	+	+
2. Textstruktur/Layout	+	–	+	o	+	o	o
3. Editionswiss. Phänomene	–	–	–	–	+	o	o
4. Linguistische Information	–	–	–	–	o	+	+
5. Kontext/Metadaten	o	–	–	o	+	–	–
6. Menschen- und maschinenlesbar	o	+	o	+	+	–	o
7. Langfristige Nutzbarkeit	o	+	–	+	+	+	+

– = nicht erfüllt, o = teils erfüllt, + = weitgehend oder voll erfüllt

T. 1: Dateiformate und Anforderungen

Um es knapp zusammenzufassen: Die Stärke von PDF-Dateien, auch wenn sie Bild- und Textinformationen beinhalten, liegt in der visuellen Präsentation des Layouts des Dokuments. Dagegen wird die Textstruktur nicht explizit gemacht und das Format erfüllt auch darüber hinaus kaum eine Anforderung. Einfache, unstrukturierte Textformate wie TXT sind zwar gut von Menschen lesbar und bleiben langfristig nutzbar, weisen aber ebenfalls eine ganze Reihe von Schwächen auf; insbesondere kann nur wenig zusätzliche Information über die Zeichensequenz hinaus, seien es Textstruktur, editorische oder linguistische Informationen oder Metadaten. Stärker strukturierte, proprietäre Formate wie DOCX haben hier ebenfalls Schwächen, vor allem aber ist ihre langfristige Nutzbarkeit zweifelhaft. Ein offenes, semi-strukturiertes Format wie XHTML kann zwar auch manche Aspekte der Textstruktur repräsentieren und ist als offener Standard langfristig nutzbar, aber die Expressivität bezüglich zusätzlicher, spezifisch literaturwissenschaftlich relevanter Informationen und Metadaten ist eingeschränkt. Formate wie XMI und TCF wiederum haben Stärken bei der Repräsentation linguistischer Annotationen, sind aber nur in geringem Maße dazu geeignet, editionswissenschaftliche Phänomene zu repräsentieren und sind kaum menschenlesbar.

Es wird deutlich, dass sich ein Format wie TEI (und in eingeschränktem Maße XHTML), also offene, nicht darstellungsorientierte, semi-strukturierte Volltextformate, für die literaturwissenschaftliche Arbeit am Besten eignen. TEI repräsentiert explizit die Zeichensequenz, die Textstruktur, lokale editionswissenschaftliche Annotationen und dokumentbezogene Metadaten. Es ist langfristig nutzbar, da es auf dem weit verbreiteten Standard XML beruht und eine große, institutionalisierte Gemeinschaft von Nutzerinnen und Nutzern hinter ihm steht. Je nach Anwendungsszenario stellen noch stärker strukturierte, nicht-proprietäre Formate wie TCF eine Alternative dar, die sich für die linguistische Annotation besser eignen, jedoch Textstruktur und Kontext weniger detailliert repräsentieren. In der Praxis wird in den Literaturwissenschaften immer noch zu häufig mit sehr einfachen Textformaten gearbeitet, die weder für die Textkonstitution ausreichend expressiv und detailreich, noch für die Textanalyse ausreichend explizit und eindeutig sind. Entweder trifft die Enttäuschung ob der begrenzten Möglichkeiten und der mangelnden Subtilität beispielsweise von Suchabfragen in DOCX-Dokumenten dann das digitale Forschungsparadigma in den Geisteswissenschaften insgesamt, oder aber die tatsächlich vorhandenen Mög-

lichkeiten digitaler Textformate für literaturwissenschaftliches Arbeiten bleiben unerkannt und damit ungenutzt. Denn die Wahl einer bestimmten Repräsentationsform ist auch jenseits technischer und praktischer Aspekte von entscheidender Bedeutung, wie Michael Sperberg-McQueen schon 1991 formuliert hat:

> Any electronic representation of a text embodies specific ideas of what is important in that text. A well-developed encoding scheme is thus in some sense a theory of the texts it is intended to mark up. […] As scholars work more intimately with computers, the electronic texts they use ought to help them in their work, making easy the kinds of work scholars want to do with them. But tools always shape the hand that wields them; technology always shapes the minds that use it. And so as we work more intimately with electronic texts, we will find ourselves doing those things that our electronic texts make easy for us to do; reason enough to think in advance about what forms electronic texts should take.[10]

LiteraturwissenschaftlerInnen sind methodisch bestens dazu in der Lage, die Konsequenzen einer gewählten Repräsentationsform zu ermitteln und sollten dies auch für digitale Repräsentationsformen tun, selbst dann, wenn sie nicht selbst (digitale) EditionswissenschaflerInnen sind. Wie viel nach den Richtlinien der Text Encoding Initiative repräsentierte Texte der literaturwissenschaftlichen Forschung sowohl für die Textkonstitution als auch für die Textanalyse zu bieten haben, aber auch inwiefern sie die Modalitäten ihrer Nutzung beeinflussen, soll in den folgenden Abschnitten aufgezeigt werden.

1.3 Die Text Encoding Initiative

Die als *Guidelines* bezeichneten Richtlinien der Text Encoding Initiative werden seit 1987 von einer internationalen, stetig wachsenden und disziplinär breit gefächerten Gemeinschaft von Wissenschaftlerinnen und Wissenschaftlern entwickelt und gepflegt. Diese Gemeinschaft hat sich in einem Konsortium institutionalisiert und verfolgt das Ziel, eine wissenschaftsadäquate Repräsentation von Texten und Dokumenten aller Art zu ermöglichen, wobei ein starker Fokus auf geisteswissenschaftlich relevanten Text- und Dokumenttypen liegt. Die TEI ist damit nicht nur als ein *de facto* Standard für die Textrepräsentation (und als einer der wenigen internationalen, geisteswissenschaftlichen Standards überhaupt) zu beschreiben, sondern auch als eine von einer Gemeinschaft von Forschenden getragene Institu-

[10] Sperberg-McQueen, „Text in the Electronic Age“, 34–5.

tion, die neben den Richtlinien selbst auch dazugehörige Ressourcen (wie Tools, Handreichungen, und eine Mailingliste) anbietet, eine Jahreskonferenz organisiert und eine Zeitschrift herausgibt.[11]

TEI wird von der Wissenschaft für die Wissenschaft entwickelt und wird laufend an weitere und neue Anforderungen angepasst. Wie die Wissenschaft selbst stetig fortschreitet, sind die Richtlinien der TEI erweiterbar und werden in der Tat ständig erweitert. Man kann als Wissenschaftler oder Wissenschaftlerin Einfluss auf die Entwicklung der Richtlinien nehmen und kann sich darauf verlassen, dass Änderungsvorschläge nach wissenschaftlichen Kriterien, nicht nach Marktfähigkeit oder Ähnlichem, beurteilt werden. Und weil dieser Anpassungsprozess schon seit 1987 kontinuierlich verfolgt wird, gibt es mittlerweile kaum noch ein editionswissenschaftlich relevantes Phänomen, das nicht in der einen oder anderen Form, meist sogar nach mehreren unterschiedlichen Strategien, in TEI repräsentiert werden könnte. Wenn Kodierungsstrategien letztlich theoretische Überzeugungen darüber ausdrücken, was Text ist und wie ein bestimmter Text funktioniert, muss der verwendete Standard flexibel sein, was die TEI ausdrücklich unterstützt. Die starke Verankerung der TEI in der Wissenschaft bedeutet auch, dass es mittlerweile nicht nur sehr viele Texte unterschiedlichster Art im TEI-Format gibt, sondern auch tausende Geisteswissenschaftlerinnen und Geisteswissenschaftler, die das Format verstehen und nutzen sowie Neueinsteigern mit Rat und Tat zur Seite stehen können.

TEI konzentriert sich auf Struktur und Bedeutung von Textelementen, weniger auf ihr Aussehen oder Layout. Genauer gesagt: TEI kodiert die Struktur des Texts und die Eigenschaften von Wörtern in Hinsicht darauf, was diese Strukturen sind oder welche Eigenschaften die Wörter haben. Wie diese Strukturen und Eigenschaften visualisiert werden sollen, ist vom jeweiligen Rezeptionskontext abhängig und kann daher separat definiert werden.[12]

[11] Für all diese Aktivitäten der TEI, siehe: www.tei-c.org. Eine sehr lesbare und frei verfügbare Überblicksdarstellung zur *Text Encoding Initiative* wurde in jüngster Zeit von Lou Burnard vorgelegt: Lou Burnard, *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources* (Marseille: OpenEdition Press, 2014), http://books.openedition.org/oep/426. In etwas knapperer Form diskutiert Fotis Jannidis Geschichte und Zukunft der TEI: Fotis Jannidis, „TEI in a Crystal Ball", *Literary and Linguistic Computing* 24, Nr. 3 (2009): 253–65, http://doi.org/10.1093/llc/fqp015. Ebenfalls lesenswert: Susan Schreibman, „Digital Scholarly Editing", in *Literary Studies in the Digital Age*, hrsg. von Kenneth M. Price und Ray Siemens (MLA, 2013), http://dlsanthology.commons.mla.org/digital-scholarly-editing.

[12] Dem gleichen Prinzip folgt die Kombination von HTML-Kodierung mit Stylesheets (CSS), wie sie im Webdesign üblich ist, oder auch die Arbeit mit Formatvorlagen in Textverarbei-

Die Repräsentation (im Sinne des Festhaltens und Speicherns) und die Präsentation (im Sinne der Visualisierung oder Darstellung) des Textes, die im Printmedium in Eins fallen, sind in der digitalen Textedition mit TEI zwei voneinander getrennte Aspekte. Dies führt zu einer erhöhten Flexibilität: zahlreiche Informationen über den Text können festgehalten und bei der Darstellung selektiv oder interaktiv visualisiert werden. Zudem kann die Visualisierung für verschiedene Kontexte in unterschiedlicher Weise erfolgen. Beispielsweise könnte ein TEI-Dokument mit vielen kodierten Ortsnamen für eine Ansicht im Browser in ein HTML-Dokument verwandelt werden, das für jeden Ortsnamen automatisch einen Link zum entsprechenden Wikipedia-Artikel oder zu einer Karte enthält. Das gleiche TEI-Dokument würde hingegen für eine gedruckte Ansicht in ein PDF-Dokument verwandelt werden, in dem die Ortsnamen lediglich für die bessere Erkennbarkeit fett gedruckt dargestellt würden.[13]

Auf einer stärker technischen Ebene, die aber ihre Bedeutsamkeit hat, ist zu erwähnen dass TEI ein prinzipiell interoperables Format ist, das unabhängig von einer bestimmten Software, Firma, oder Plattform ist. Auch wenn eine TEI-Datei auf einem Windows-Rechner erstellt wurde, ist sie auf einem Macintosh oder einem Linux-Rechner problemlos verwendbar.[14] Und weil TEI technisch gesehen wie bereits erwähnt XML ist, d.h. den Grundprinzipien eines extrem weit verbreiteten Datenformats folgt, kann es auch von einer äußerst großen Anzahl von sehr generischen, weit verbreiteten Tools und mit allen für XML entwickelten Techniken weiter bearbeitet werden. Einige weitere, grundlegende Eigenschaften von XML, die unabhängig von TEI sind und von XML gewissermaßen an TEI vererbt werden, sind noch erwähnenswert: die Möglichkeit, einige mit XML verwandte Technologien zu verwenden;[15] die Möglichkeit, ein XML-kodiertes Dokument auf seine Wohlge-

tungsprogrammen wie Word oder LibreOffice; TEI setzt die Trennung von Eigenschaften und Visualisierung jedoch noch konsequenter um.

[13] Für weitere Beispiele, siehe Christof Schöch und Johanna Wolf, „Die Visualisierung von Varianten in der Textedition: alte methodische Debatten im neuen Licht der digitalen Medien“, in *Variante und Varietät: Akten des VI. Dies Romanicus Turicensis*, hrsg. von Cristina Albizu u.a. (Pisa: Edizioni ETS, 2013), 189–206, https://hal.archives-ouvertes.fr/hal-00945358. Dieses sogenannte „single-source publishing“ hat sich u.a. auch im Bereich von Webseiten etabliert, die in Abhängigkeit der Bildschirmgröße unterschiedlich detaillierte Darstellungen anbieten.

[14] Dies setzt lediglich voraus, dass für die Zeichenkodierung UNICODE/UTF-8 verwendet wurde.

[15] Diese Technologien, zu denen XPath, XQuery und XSLT gehören, werden hier nicht disku-

formtheit zu überprüfen, d.h. auf die Einhaltung der Regeln von XML; und die Möglichkeit, ein XML-kodiertes Dokument auf seine Validität bezüglich eines bestimmten Schemas zu überprüfen. Nur auf das letzte dieser Themen, die Validierung, sei hier noch knapp eingegangen: Im Kern bedeutet dies, dass ein TEI-Dokument daraufhin überprüft werden kann, ob es den für ein Kodierungsprojekt als relevant erachteten Teil der Richtlinien der TEI konsequent und einheitlich respektiert. Dies bedeutet nicht nur, dass die editorische Arbeit zumindest teilweise auf (formale, nicht aber inhaltliche) Konsistenz überprüft werden kann, sondern auch, dass eine menschen- und maschinenlesbare Zusammenfassung der Kodierungsstrategie für ein Projekt vorliegt, was das Verständnis und die langfristige Nutzbarkeit einer Textedition erhöht. Viele Programme zur Bearbeitung von XML-Dateien unterstützen diese Art von Kohärenzprüfung.

Neben den zahlreichen Vorzügen der Repräsentation von Texten nach den Richtlinien der TEI sollten auch einige Aspekte nicht verschwiegen werden, unter denen TEI die Möglichkeiten der Textrepräsentation auf teilweise problematische Art einschränkt oder zumindest lenkt. Ein erster Aspekt betrifft den Umstand, dass die Richtlinien notwendiger Weise auf diejenigen Eigenschaften von Texten fokussiert ist, die in einem bestimmten Dokumenttyp immer wieder in erkennbar ähnlicher Weise vorkommen, seien es nun Strukturmerkmale wie Kapitel oder Absätze oder lokale Phänomene wie Personennamen oder das Auftreten von Metaphern. Ein anderes Vorgehen würde nicht nur die Idee eines von einer breiten Gemeinschaft geteilten Standards *ad absurdum* führen, es würde auch dem Prinzip der Vergleichbarkeit und Kategorisierbarkeit von Einzelphänomenen widersprechen. Zwar ist demnach festzuhalten, dass die letztlich einzigartigen Aspekte eines bestimmten Textes nicht im Fokus der TEI stehen; allerdings sind die bereits vorhandenen Möglichkeiten von einer außerordentlichen Detailfülle, Nuanciertheit und Präzision, die nicht unterschätzt werden sollten. Für die literaturwissenschaftliche Analyse ebenfalls interessante, sehr spezielle Eigenschaften von bestimmten Einzeltexten oder individuelle Abweichungen von einem regelhaft beschreibbaren System, sind nicht unmittelbar, wohl aber indirekt (bspw. durch individuelle, beschreibende Attribute oder andere, spezielle Mechanismen) repräsentierbar. Ein kaum vermeidbarer Nachteil von TEI ist, dass linguistische und literaturwissenschaftlich orien-

tiert; stattdessen sei auf die folgende, einführende Darstellung verwiesen: Helmut Erlenkötter, XML – *Extensible Markup Language von Anfang an* (Reinbek bei Hamburg: Rowohlt, ²2008).

tierte Annotationen auf Wortebene zwar grundsätzlich möglich sind, aber insbesondere in Kombination mit textkritischen Auszeichnungen zu einer gewissen Unübersichtlichkeit führen und den Prozessierungsaufwand erhöhen können. Hier sind primär für einen solchen Zweck entwickelte Formate (wie die in Abschnitt 2 genannten) der TEI überlegen.

Trotz dieser zuletzt genannten Aspekte ist TEI insgesamt eindeutig das Mittel der Wahl für die Repräsentation von Texten, die für literaturwissenschaftliche Arbeit eingesetzt werden sollen. Einige wenige der Möglichkeiten zur systematischen Repräsentation von Texten, ihrer Struktur und ihres Kontextes, wie sie die Richtlinien der Text Encoding Initiative anbieten, werden daher im nächsten Abschnitt vorgestellt, bevor dann auch auf die Nutzung der TEI für die Textanalyse eingegangen wird.

2. Möglichkeiten der TEI bei der Textkonstitution: Prinzipien und Anwendungsbeispiele

Zunächst einmal ist festzuhalten, dass TEI eine spezifische Form von XML ist, also eine Auszeichnungssprache. Daraus ergibt sich zunächst ganz allgemein, dass in TEI repräsentierte Texte einerseits die Zeichenkette repräsentieren, andererseits weitere Informationen über die Zeichenkette direkt im Text mit markiert, d.h. ausgezeichnet werden. Dabei werden die Auszeichnungen durch spitze Klammern vom Text selbst unterschieden. Ein Einzelsatz in TEI, der lediglich die Satzgrenzen (mit dem öffnenden `<s>` und schließenden `</s>`-Tag) und einen Ortsnamen (als `<placeName>`) auszeichnet, könnte also folgendermaßen aussehen wie in Beispiel 2.

```
<s>Como todos los hombres de <placeName>Babilonia</placeName>,
he sido procónsul; como todos, esclavo.</s>
```

T. 2: Kodierung eines Satzes mit Ortsnamen

Die Auszeichnungen in spitzen Klammern werden Tags genannt; es gibt immer ein öffnendes und ein schließendes Tag. Die beiden zusammengehörigen Tags und ihr Textinhalt werden Element genannt. Ein Element kann auch Attribute mit bestimmten Werten haben, die das Element oder den Textinhalt weiter spezifizieren, sowie weitere Elemente enthalten. Im Beispiel 3 wurde mit einer Attribut-Wert-Kombination festgehalten, dass sich der Ortsname auf eine Stadt bezieht.

```
<s>Como todos los hombres de <placeName type="Stadt">Babilonia</placeName>,
he sido procónsul; como todos, esclavo.</s>
```

T. 3: Kodierung eines Satzes mit Attributen/Werten

Nach diesem Prinzip können im Textverlauf zahlreiche Informationen festgehalten und hunderte Phänomene charakterisiert werden. Lou Burnard beschreibt prägnant, wie sich die Richtlinien der TEI in ihrer Gesamtheit zu einem bestimmten Einzelprojekt verhalten:

> The TEI provides names and definitions for many hundred tags, together with rules about how they may be combined. More exactly, the TEI Guidelines define some five or six hundred different concepts, along with detailed specifications for the XML elements and element classes which may be used to represent them. Most, if not all, TEI documents need to use only a small amount of what is provided.[16]

In der Tat sind die Richtlinien der TEI inzwischen zu einem Kompendium textwissenschaftlich relevanter Phänomene geworden, das diese definiert, beschreibt, voneinander abgrenzt und an Beispielen illustriert und hohen Informationswert hat. Die TEI stellt Gruppen von Elementen (sogenannte Module) für Prosa, Lyrik und dramatische Texte bereit, außerdem spezielle Module unter anderem für Manuskripte, kritische Apparate, Wörterbücher, gesprochene Sprache und linguistische Korpora. Einzelprojekte werden demnach immer nur einen kleinen Teil der Richtlinien tatsächlich benötigen. Zu den grundlegenden Konzepten, die in TEI repräsentiert werden können, gehören insbesondere Informationen über die Textstruktur, Annotationen auf lexikalischer Ebene, texteditorische Informationen und Informationen über das Dokument.

In den Bereich der Informationen über die Textstruktur fällt die Auszeichnung der Makrostruktur eines Textes. Beispielsweise kann sich ein Roman in Vorworte und Haupttext gliedern, der Haupttext wiederum in Teile, Kapitel und Absätze. Ein Theaterstück kann sich in Vorwort, Prolog und Haupttext, der Haupttext wiederum in Akte, Szenen, Bühnenanweisungen und Reden gliedern, die Reden wiederum Sprechernamen und Sprechertext enthalten. Die TEI bietet für diese Struktureinheiten Elemente an: Unter anderem wird der Haupttext von voranstehendem und folgendem Peritext un-

[16] Burnard, *What is the TEI?*, 16. Die Richtlinien selbst sind unter folgendem Link einsehbar: www.tei-c.org/release/doc/tei-p5-doc/en/html.

terschieden, wobei das Element `<body>` den Haupttext beinhaltet, `<front>` u.a. für Vorworte oder Widmungen und `<back>` u.a. für Indices, Glossare oder Nachworte verwendet wird. Innerhalb dieser Elemente sind verschiedene Abschnitte mit ihren Titeln (in `<div>` und `<head>`) vorgesehen. Innerhalb von Abschnitten wiederum können Absätze und Sätze (in `<p>` und `<s>`), Verszeilengruppen und Verszeilen (in `<lg>` und `<l>`) oder Figurenreden (in `<sp>` für „speech“) mit Sprechernamen (in `<speaker>`) und der Rede selbst (in `<p>` oder `<l>` bzw. `<lg>`) vorkommen. Beispiel 4 gibt einen kurzen Abschnitt aus Heinrich von Kleists *Zerbrochnem Krug* in TEI-Format wieder.

```
<body>
    <div type=”auftritt”>
        <head n=”1”>Erster Auftritt</head>
        <stage>Adam sitzt und verbindet sich ein Bein. Licht tritt auf.</stage>
        <sp>
            <speaker>LICHT.</speaker>
            <lg>
                <l>Ei, was zum Henker, sagt, Gevatter Adam!</l>
                <l>Was ist mit Euch geschehn? Wie seht Ihr aus?</l>
            </lg>
        </sp>
        <sp>
            <speaker>ADAM.</speaker>
            <lg>
                <l>Ja, seht. Zum Straucheln braucht’s doch nichts, als Füße.</l>
                <l>Auf diesem glatten Boden, ist ein Strauch hier?</l>
                <l>Gestrauchelt bin ich hier; denn jeder trägt</l>
                <l>Den leid’gen Stein zum Anstoß in sich selbst.</l>
            </lg>
        </sp>
    </div>
</body>
```

T. 4: Kodierung von Theaterstücken

Innerhalb des Textes können Einzelwörter oder Wortfolgen ihren spezifischen Eigenschaften oder ihrer Bedeutung nach beschrieben werden, beispielsweise als Personenname (Element `<persName>`), als Ortsname (mit `<pla-`

ceName>) oder als Zeitausdruck (mit <time> oder <date>). Eine andere Gruppe von Elementen dient dazu, Wörter beispielsweise als Werktitel (also <title>), fremdsprachige Ausdrücke (mit <foreign>) oder emphatisch verwendete Begriffe (mit <emph>) auszuzeichnen. (Das Prinzip wurde oben bereits illustriert.) Hier wird deutlich, dass das Grundprinzip, nicht Aussehen, sondern Bedeutung oder Eigenschaften eines Worts zu markieren, Vorteile hat: Sowohl Titel, fremdsprachige Ausdrücke und emphatischer Wortgebrauch werden üblicherweise im Text gleichermaßen durch Kursivierung visualisiert und insofern nicht ausdrücklich unterschieden. Sie jeweils explizit zu benennen bedeutet allerdings auch, dass jeweils eine editorische Entscheidung und ggfs. Interpretation notwendig ist.

Bezüglich der textkritischen Informationen kann hier nicht einmal ansatzweise auf die äußerst umfangreichen und detaillierten Mechanismen eingegangen werden, die hierfür zur Verfügung stehen. Am Beispiel der punktuellen Korrektur oder Modernisierung von Texten soll nur in aller Kürze aufgezeigt werden, wie flexibel und detailliert die TEI hier ist. Das vereinfachte Beispiel 5 stammt aus der digitalen Edition des *Essai sur le récit* von Bérardier de Bataut. Der ursprünglich 1776 erschienene Text weist sowohl historische Grafien als auch orthographische Fehler auf, mit denen auf unterschiedliche Weise umgegangen werden kann. Erstens können solche Phänomene schlicht als solche markiert werden, ohne dass in den Text eingegriffen würde.

```
<p>Ce témoignage ici ne <sic>peut-être</sic> suspect. C’est
celui d’un <orig>Auteur</orig> de romans.</p>
```

T. 5: Verwendung von <sic> und <orig>

Die Elemente <sic> und <orig> sind jeweils dazu gedacht, offensichtliche Fehler bzw. historische Grafien u.ä. zu markieren. Möchte man statt der originalgetreuen Transkription einen korrigierten und modernisierten Text anbieten, kann diese editorische Intervention dokumentiert werden, vgl. Bsp. 6.

```
<p>Ce témoignage ici ne <corr>peut être</corr> suspect. C’est
celui d’un <reg>auteur</reg> de romans.</p>
```

T. 6: Verwendung von <corr> und <reg>

Mit Hilfe der Elemente `<corr>` und `<reg>` (für Korrekturen und Normalisierungen) kann nun nachvollzogen werden, an welchen Stellen vom Editor eine Veränderung gegenüber der Vorlage vorgenommen wurde. Um den Lesern genauer aufzuzeigen, worin die Veränderung genau liegt und wie die ursprüngliche Fassung lautete, können auch beide Varianten als Alternativen festgehalten werden, die durch das Element `<choice>` verbunden werden, vgl. Bsp. 7.

```
<p>Ce témoignage ici ne
    <choice>
        <sic>peut-être</sic>
        <corr>peut être</corr>
    </choice> suspect. C’est celui d’un
    <choice>
        <orig>Auteur</orig>
        <reg>auteur</reg>
    </choice> de romans.</p>
```

T. 7: Verwendung von `<choice>`

Für die Darstellung des Textes kann nun anschließend die eine oder andere Fassung gewählt werden, oder den LeserInnen interaktiv die Möglichkeit gegeben werden, selbst die jeweilige Fassung auszuwählen. Dies ist beispielsweise in der digitalen Edition des *Essai sur le récit* von 2010 umgesetzt, wie Abbildung 1 zeigt.[17]

Schließlich können mit Attributen weitere Informationen zu diesen Varianten festgehalten werden, unter anderem: Um welche Art von Fehler handelt es sich? Wer im Herausgeberteam oder in der Forschung hat diesen Fehler festgestellt? Wie sicher ist sich diese Person, dass es sich in der Tat um einen Fehler handelt? Hierfür stehen die Attribute `type`, `resp` (responsibility) und `cert` (certainty) zur Verfügung, deren Verwendung das Beispiel 8 an nur einem Element illustriert.

Es sei angemerkt, dass für texteditorische Informationen viele Dutzend weitere Elemente und Attribute verfügbar sind, die sich sowohl für klassi-

[17] Siehe François-Joseph Bérardier de Bataut, *Essai sur le récit, ou Entretiens sur la manière de raconter* (Paris : Charles-Pierre Berton, 1776). Édition électronique sous la direction de Christof Schöch, 2010, www.berardier.org.

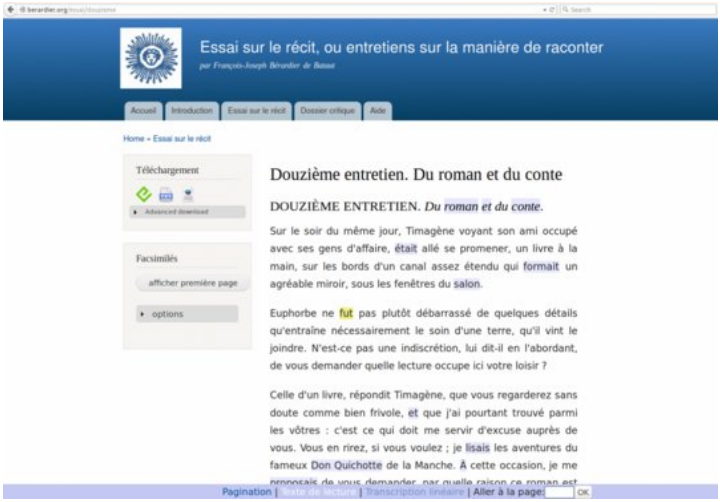

Abb. 1: Screenshot der digitalen Edition des *Essai sur le récit*, http://berardier.org, 2010.

```
<sic type="Interpunktion" resp="Delon-2011" cert="mittel">peut-être</sic>
```

T. 8: Verwendung von Attributen

sche kritische Editionen mit Apparat, als auch für genetische Texteditionen oder auch für Editionen mit dokumentzentrierter Perspektive eignen.[18]

In den ersten Jahrzehnten seiner Existenz war TEI in erster Linie auf den Text als solchen und auf seine hierarchische Struktur fokussiert, nicht auf die räumliche Anordnung und den materiellen Träger des Textes. Dies hat sich für bestimmte aktuelle Forschungsfragen aus dem Bereich der Bild/Text-Beziehungen und der *critique génétique* als hinderlich erwiesen. In neuerer Zeit hat die Community in diesem Bereich reagiert und unter anderem das TEI-Modul „Representation of Primary Sources" mit Blick auf die dokumentzentrierte Edition weiterentwickelt.[19] Das Modul erlaubt es nun beispielsweise, Informationen über die Eigenschaften des materiel-

[18] An weiterführenden Informationen interessierte Leser seien auf die *Guidelines* der TEI selbst verwiesen, aber auch auf hierauf spezialisierte Einführungstexte: Lou Burnard, Katherine O'Brien O'Keeffe und John Unsworth, Hrsg., *Electronic Textual Editing* (New York: MLA, 2006).

[19] Es handelt sich hier nebenbei um ein ausgezeichnetes Beispiel dafür, wie die Richtlinien der TEI auf Bedürfnisse der Community angepasst werden können, wenn eine gut koordinierte Initiative Vorschläge für Erweiterungen macht.

len Textträgers und über die räumliche Gliederung des Textes auf der Seite festzuhalten. Das Element `<surface>` definiert die Gesamtfläche eines texttragenden Objektes, `<zone>` beschreibt beliebige Bereiche innerhalb der Gesamtfläche, und `<line>` bezeichnet typographische Zeilen. Bestimmte Bereiche, beispielsweise einzelne Textblöcke, Marginalien oder einzelne Textzeilen können über räumliche Koordinaten in einem digitalen Faksimile präzise lokalisiert und mit dem transkribierten Text verknüpft werden.

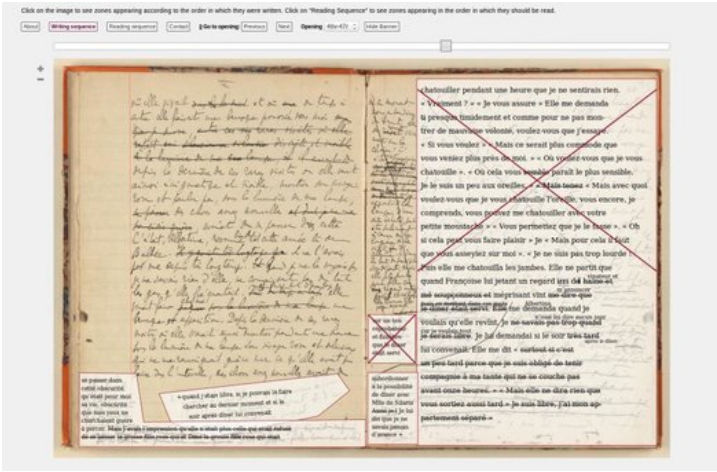

Abb. 2: Zonen und Schreibprozess am Beispiel eines Proust-Manuskripts, aus: *Autour d'une séquence et des notes du Cahier 46*

In Kombination mit erweiterten Möglichkeiten für die Darstellung von Textveränderungen und der Möglichkeit, diese verschiedenen Überarbeitungsphasen zuzuordnen, ergeben sich neue Möglichkeiten für die Dokumentation und Visualisierung von Schreibprozessen. Die strukturorientierte Ausrichtung der TEI erfährt auf diese Weise eine Ergänzung um die räumliche und zeitliche Dimensionen. Abbildung 2 zeigt einen Screenshot aus einer interaktiven Anwendung dieses neuen Paradigmas am Beispiel eines Proust-Manuskripts.[20]

Ein weiteres Merkmal von TEI-Dokumenten ist die Tatsache, dass sie sowohl den zu repräsentierenden Text selbst, als auch eine ganze Reihe von In-

[20] Abbildung 2 zeigt einen Screenshot aus *Autour d'une séquence et des notes du Cahier 46: enjeu du codage dans les brouillons de Proust*, hg. von Elena Pierazzo und Julie André, 2011, http://research.cch.kcl.ac.uk/proust_prototype/about.html. Bildlizenz: Creative-Commons Attribution Non-Commercial.

formationen über den Text (sogenannte Metadaten) in einem Dokument beinhalten. Die Bedeutung solcher Informationen für die langfristige Nutzbarkeit eines Dokuments in verschiedenen Kontexten ist kaum zu überschätzen, weil sie wesentliche Eigenschaften des Dokuments explizit machen und Kontextinformationen mit überliefern. Erneut gilt auch hier, dass insbesondere maschinenlesbare, standardisierte Metadaten besonders wertvoll sind. Da sich diese Informationen auf das gesamte Dokument beziehen, können Sie in einem eigenen Bereich zusammengefasst werden. Entsprechend besteht ein TEI-Dokument grundsätzlich aus einem sogenannten `<teiHeader>` (für die Metadaten) und einem `<text>`-Element (für den eigentlichen Textinhalt). Alle bisher besprochenen Phänomene und Elemente kommen in der Regel im `<text>` vor. Die folgenden Ausführungen beziehen sich auf den `<teiHeader>` (der in einem TEI-Dokument allerdings vor dem Haupttext platziert ist). Der teiHeader kann, vereinfacht gesagt, unter anderem die folgenden Arten von Informationen über den Text in mehr oder weniger strukturierter Form beinhalten:

- Um welchen Text handelt es sich? Hier werden unter anderem Autor, Titel und Herausgeber genannt, oder auch ein eindeutiger Identifier für Autor und Text eingebracht (im `<titleStmt>`).
- Auf welcher Quelle beruht der Text? Hier kann die dem digitalen Text zugrunde liegende Printausgabe oder andere Quelle genannt und beschrieben werden (in der `<sourceDesc>`).
- In welcher Form und unter welchen Bedingungen ist der Text verfügbar? Hier kann festgehalten werden, ob der Text urheberrechtlichen Beschränkungen unterliegt und nach welcher Lizenz er veröffentlicht wurde (Element `<publicationStmt>`).
- Nach welchen Prinzipien wurde der digitale Text erstellt? In diesem Teil kann beispielsweise dokumentiert werden, ob und wenn ja nach welchen Regeln der digitale Text normalisiert, modernisiert, strukturiert und annotiert wurde (in der `<encodingDesc>`).
- Welche grundlegenden Eigenschaften hat das Werk? Hier werden unter anderem das Datum der Erstpublikation, der Erscheinungsort, die Gattungszugehörigkeit oder auch die Länge und Sprache des Werks genannt (in der `<profileDesc>`).

Beispiel 9 illustriert einen `<teiHeader>` für ein spanisches Theaterstück, der einen Teil der genannten Informationen enthält.

```
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title type="main">Examen de maridos</title>
            <title type="short">Examen</title>
            <author>
                <name type="full">Juan Ruiz de Alarcón</name>
                <name type="short">Alarcon</name>
                <idno type="VIAF">88975006</idno>
            </author>
            <editor xml:id="js">Jakob Stahl</editor>
        </titleStmt>
        <publicationStmt>
            <publisher>CLiGS</publisher>
            <availability status="publicdomain"/>
            <date>2015</date>
            <idno type="cligs">te0123</idno>
        </pulicationStmt>
        <sourceDesc>
            <bibl type="digital-source">
                <ref target="www.comedias.org"/>, <date>1998</date>.
            </bibl>
            <bibl type="print-source">
            Edición princeps, Secunda parte de <title>Las comedias de Don
            Juan Ruiz de Alarcón</title>, Barcelona, <date>1634</date>.
            </bibl>
        </sourceDesc>
    </fileDesc>
    <profileDesc>
        <textClass>
            <keywords>
                <term type="genre">comedy</term>
                <term type="genrelabel">comedia</term>
                <term type="form">verse</term>
            </keywords>
        </textClass>
    </profileDesc>
</teiHeader>
```

T. 9: Der teiHeader

An den beiden obigen, etwas längeren Beispielen lässt sich eine wichtige Eigenschaft von XML und damit von TEI-Dokumenten erkennen, nämlich die Tatsache, dass sie prinzipiell eine streng hierarchische Struktur haben. Diese kann durch die Einrückung sichtbar gemacht werden. Im obigen Header enthält das `<teiHeader>`-Element ein `<fileDesc>`-Element; dieses wiederum ein `<titleStmt>`-Element; dieses wiederum mehrere `<title>`-Elemente und ein `<author>`-Element, etc. Das zugrundeliegende Textmodell, das TEI von XML erbt, wird OHCO (Ordered Hierarchy of Content Objects) genannt und ist natürlich nur eines unter vielen möglichen Textmodellen.[21] Das hierarchische Textmodell sieht beispielsweise nicht vor, dass ein Satz über Absatzgrenzen hinweg verläuft; ein solches und weitere Phänomene überlappender Hierarchien, die in literarischen Texten durchaus vorkommen, können allerdings ebenfalls kodiert werden. Ein triviales Beispiel hierfür ist die Kodierung von Zeilenumbrüchen in einer bestimmten gedruckten Ausgabe, die natürlich meist nicht mit Satz- oder Absatzgrenzen übereinstimmen. Sie können mit einem sogenannten Milestone-Element (in diesem Fall `<lb/>` für *line break*) markiert werden, das selbst keinen Inhalt hat, sondern eine bestimmte Position im Text markiert. Das heißt, Druckzeilen werden nicht als Textabschnitt mit Anfang und Ende markiert, sondern es werden die Zeilenumbrüche markiert.

Während im Textbeispiel 10 aus Maupassants *Une Vie* die Satzstruktur mit öffnenden und schließenden Tags (also mit `<s>` und `</s>`) markiert wird, sind die typographisch relevanten Zeilenumbrüche jeweils mit einem leeren `<lb/>`-Element markiert.[22] Eine bestimmte typographische Zeile reicht dann von einem dieser Elemente zum nächsten. Das Attribut `ed` dient hier dazu, in verkürzter Form anzugeben, welcher Ausgabe diese Layoutinformation entspricht. Es könnten demnach auch die abweichenden Zeilenumbrüche mehrere Ausgaben kodiert werden. Weiterführende Informationen über die referenzierte(n) Ausgabe(n) können im `teiHeader` angeboten werden. Was hier aus Gründen der Darstellung an Druckzeilen illustriert wurde, gilt ebenso natürlich auch für Seitenzahlen.

Zusammenfassen lässt sich demnach festhalten, dass TEI ein reichhaltiger und flexibler *de facto* Standard für die Textkonstitution und Textrepräsentation ist. Die TEI ermöglicht die Erstellung wissenschaftlicher Editionen lite-

[21] Zu OHCO, siehe Steven DeRose u. a., „What Is Text, Really?“ *Journal of Computing in Higher Education* 1, Nr. 2 (1990): 3–26.

[22] Der für die Darstellung des Beispiels hier gewählte, implizite Zeilenumbruch hat dagegen keine Bedeutung für das Dokument.

```
<div>
  <p>
    <s>Jeanne, ayant fini ses malles, s'approcha de la fenêtre,
    mais <lb ed="1975" /> la pluie ne cessait pas.</s>
  </p>
  <p>
    <s>L'averse, toute la nuit, avait sonné contre les carreaux
    et <lb ed="1975" />les toits.</s>
    <s>Le ciel, bas et chargé d'eau, semblait crevé, se
    <lb ed="1975" /> vidant sur la terre, la délayant en bouillie,
    la fondant comme <lb ed="1975" />du sucre.</s>
    <s>Des rafales passaient, pleines d'une chaleur lourde. <lb ed="1975" /></s>
    <s>Le ronflement des ruisseaux débordés emplissait les rues
    désertes <lb ed="1975" /> où les maisons, comme des éponges,
    buvaient l'humidité qui pénétrait <lb ed="1975" />au-dedans et
    faisait suer les murs de la cave au grenier.</s>
  </p>
</div>
```

T. 10: Textbeispiel aus Maupassant, *Une Vie*

rarischer Texte, die in ihrer Präzision und Zuverlässigkeit gedruckten Ausgaben in nichts nachstehen müssen und diesen durch ihre vielseitige Nutzbarkeit klar überlegen sind. (Eine Reihe von Webportalen, die Anlaufstellen für digitale Texte im TEI-Format sind, werden im Anhang zu diesem Beitrag genannt.) Ein entscheidender Mehrwert ist, dass die so sorgfältig und an einer Stelle festgehaltenen Informationen in vielfältiger Form visualisiert und präsentiert werden können: in unterschiedlichen Formaten (bspw. in HTML für die Ansicht am Bildschirm, als EPUB für die Lektüre auf dem E-Reader, und als PDF für den Druck), aber auch für unterschiedliche Zielgruppen (bspw. mit modernisiertem Text und Sachkommentaren für jüngere LeserInnen, mit Varianten und textkritischem Apparat für professionelle LeserInnen). Und nicht nur unterschiedliche Ansichten des Textes für unterschiedliche Lektüresituationen werden möglich, sondern die Informationen stehen darüber hinaus auch in maschinenlesbarer Form für qualitative und quantitative Analysen zur Verfügung. Dies gilt sowohl bei Texten geringerer Zahl und überschaubaren Umfangs als auch dann, wenn es um Textmengen geht, die

nicht mehr ohne Weiteres überschaubar sind. Einige Möglichkeiten solcher Analysen sollen im folgenden Abschnitt aufgezeigt werden.

3. Möglichkeiten der TEI bei der Textanalyse: Prinzipien und Anwendungsbeispiele

Die Verwendung von TEI-kodierten Texten für die computergestützte, quantitative Textanalyse ist eine relativ neue Entwicklung, die durch die zunehmende Anwendung quantitativer Verfahren auf literarische Texte befördert wird. Grundsätzlich gilt, dass insbesondere die (maschinenlesbare) Auszeichnung von Strukturmerkmalen von Texten es bei der Textanalyse erlaubt, diese Information gezielt zu nutzen. Dadurch können Suchabfragen präziser gestellt werden und Vereinfachungen vermieden werden. Aber auch andere, in TEI-kodierten Texten vorliegende Informationen können für die literaturwissenschaftliche Analyse und Interpretation interessant sein, wie die folgenden Abschnitte zeigen möchten. Der digital vorliegende, strukturierte Text erleichtert die Bearbeitung etablierter Fragestellungen, ermöglicht aber auch ganz neue methodische Zugriffe auf Texte.[23]

3.1 Suche nach Personen, Orten oder Begriffen

In vielen Fällen richtet sich das Interesse einer literaturwissenschaftlichen Analyse auf bestimmte Elemente der fiktionalen Welt, wie beispielsweise Figuren oder Orte, oder auf bestimmte Referenzen, wie die auf andere Autoren und deren Werke, oder auf bestimmte Wörter, wie beispielsweise ein Konzept oder einen bestimmten Begriff.

Eine einfache Suche in einem unstrukturierten Textdokument fördert dabei häufig Treffer zu Tage, die eigentlich nicht relevant sind oder zumindest unterschieden werden sollten. Suchen wir beispielsweise in einer größeren Textsammlung nach der Zeichenkette „Madame Bovary", werden wir in der Trefferliste sowohl Passagen finden, in denen die fiktionale Figur erwähnt wird, als auch Passagen, in denen Flauberts Romantitel erwähnt wird. Die Suche nach einem bestimmten Begriff in einem Theaterstück wird alle Treffer im gesamten Text zu Tage fördern, auch wenn das Interesse der Analyse

[23] Für eine einführende Darstellung, siehe Fotis Jannidis, „Methoden der computergestützten Textanalyse", in *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*, hrsg. von Ansgar Nünning and Vera Nünning (Stuttgart & Weimar: Metzler, 2010), 109–32. Für eine Reflexion über die methodischen Konsequenzen des digitalen Textes für die literaturwissenschaftliche Analyse und Interpretation, siehe auch Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (Urbana Ill.: University of Illinois Press, 2011).

sich nur auf die Verwendung des Begriffs durch eine spezifische Figur richtet. Und die Suche nach Orten könnte daran scheitern, dass man vielleicht gar nicht weiß, welche Orte in einem Text vorkommen, sondern überhaupt erst einmal eine Liste aller Ortsnamen erstellen möchte.

In einem nach den Richtlinien der TEI kodierten Dokument können diese und viele weitere Verwendungsweisen und Kontexte einer Zeichenkette klar unterschieden und ausgezeichnet werden. Bei der Zeichenkette „Madame Bovary" handelt es sich um einen Buchtitel (kodierbar mit dem Element `<title>`) oder aber einen Personennamen (Element `<persName>`). Eine (beispielsweise mit einem sogenannten XPath formulierte) Suchabfrage in TEI-Dokumenten ist in der Lage, diese Information zu nutzen und bspw. nur diejenigen Zeichenketten „Madame Bovary" als Treffer anzuzeigen, die innerhalb des Elements `<title>` erscheinen. Ebenso wird in einem Theaterstück jede Figurenrede mit dem Sprecher verbunden, der für sie verantwortlich ist. Dadurch können bei der Suche nach einem Begriff die Treffer entweder auf die Reden einer oder mehrerer Figuren eingegrenzt, oder aber durch die Information über die dazugehörige Figur ergänzt werden.

Schließlich kann in einem in geeigneter Weise kodierten TEI-Dokument auch nach allen Zeichenfolgen gesucht werden, die sich in einem `<placeName>`-Element befinden. Ausgehend von solchen Daten können dann die erwähnten Orte auch mit geographischen Koordinaten versehen und/oder auf einer aktuellen oder historischen Karte angezeigt werden. Ein Beispiel für ein solches Vorgehen ist das *Map of Early Modern London*-Projekt, das eine historische London-Karte von 1561 mit Zitaten aus zahlreichen zeitgenössischen Werken verbindet, in denen bestimmte Straßennamen und andere Orte in London vorkommen.[24] Das Beispiel zeigt, wie durch eine derartige Kodierung und Visualisierung ein neuer, räumlich organisierter Blick auf literarische und nicht-literarische Texte enstehen kann.

Abbildung 3 zeigt einen Ausschnitt aus der interaktiven *Map of Early London*.[25] Hier wurde eine bestimmte Kirche (St. Alban Church, Wood Street) gelb hervorgehoben. Rechts erscheint eine Liste der Texte, in denen diese Kirche erwähnt wird.

[24] Map of Early Modern London-Projekt, siehe: http://mapoflondon.uvic.ca.

[25] Bildquelle: Screenshot aus dem *Map of Early Modern London*-Projekt, http://mapoflondon.uvic.ca, Creative Commons Attribution Share-Alike 4.0.

Abb. 3: Ansicht aus dem Map of Early London Project (Screenshot)

3.2 Kookkurrenzanalyse

Für die Nutzung von TEI-Auszeichnungen für die Textanalyse ist es nicht immer notwendig, die entsprechenden Suchabfragen selbst mit Hilfe von Technologien wie XPath zu formulieren. Es gibt auch benutzerfreundliche Software, die die Strukturinformationen in TEI-Dokumenten direkt ausnutzt. Ein solches Tool ist TXM, das den Import von TEI-kodierten Texten erlaubt.[26] Auch einfache TEI-Dokumente, die beispielsweise lediglich Informationen über Kapitel, Absätze und Sätze in TEI-kodierter Form beinhalten, können für bestimmte Suchabfragen sinnvoll eingesetzt werden. Ein Beispiel für eine solche Analyse, die von TXM unterstützt wird, ist die Kookkurrenz-Analyse. Die grundlegende Frage ist hier, welche Wörter in einem bestimmten Text ungewöhnlich häufig mit einem bestimmten anderen Begriff gemeinsam auftreten. Solche Analysen können beispielsweise Aufschluss darüber geben, wie ein bestimmter Begriff in einem Text bewertet wird.

Üblicherweise definiert man, um den Sachverhalt des „gemeinsamen Auftretens“ genauer zu bestimmen, ein „Fenster“ rund um das Zielwort, in dem nach anderen Wörtern gesucht werden soll. Als „gemeinsam auftretend“ gelten dann bspw. nur die Wörter, die höchstens zehn Wörter vor und zehn

[26] Zu TXM, siehe http://textometrie.ens-lyon.fr.

Wörter nach dem Zielwort auftreten. Nur diese gehen in die Berechnung der Häufigkeiten des gemeinsamen Auftretens ein. Dies erlaubt eine präzise Eingrenzung des Zielfensters, zugleich kann dies allerdings auch zu Ungenauigkeiten führen. Denn innerhalb dieses Fensters können ja auch Satz- oder Absatzgrenzen liegen. Es ist aber nun nicht gleich zu bewerten, ob ein bestimmtes Wort im gleichen Satz wie das Zielwort vorkommt, oder aber zwar auch in geringem Abstand, aber in einem anderen Satz oder Absatz.

Genau diese Tatsache berücksichtigt nun TXM, das die TEI-Auszeichnung für diese Art von Analyse nutzen kann. Beispielsweise ist es möglich, das Fenster um das Zielwort eben nicht als Anzahl von Wörtern, sondern als (dann wesentlich kleiner zu wählende) Anzahl von Sätzen vor oder nach dem Zielwort zu definieren. Somit ist es dann eben auch möglich, nur solche Wörter als „gemeinsam auftretend" aufzufassen, die im gleichen Satz wie das Zielwort auftreten, unabhängig von der Länge der Sätze. Ein solches Suchverfahren erhöht nicht nur die Genauigkeit der Analyse, es ist auch philologisch präziser.

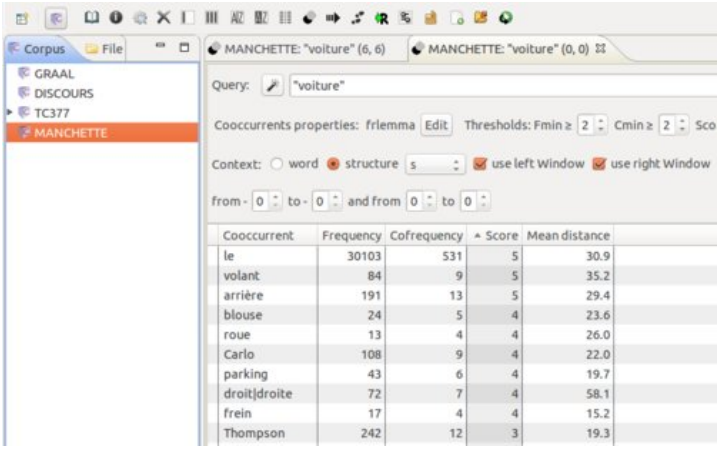

Abb. 4: Kookkurrenz-Analyse in TXM (Screenshot)

Abbildung 4 zeigt eine entsprechende Suchabfrage in TXM, in der eine Sammlung von Kriminalromanen Jean-Patrick Manchettes, die in TEI vorliegen, untersucht wurde. Die Abfrage zeigt die Wörter, die statistisch gesehen unerwartet häufig im gleichen Satz wie „voiture" auftreten: neben dem Artikel („la voiture", hier wird das Lemma „le" angezeigt) sind dies andere Teile des Autos oder spezifische, zum Autofahren gehörige Gegenstände. In der

Mitte des Screenshots sieht man, dass das Strukturmerkmal „s“ (für Satz) verwendet wurde, und im Abstand von „0“ Sätzen vor und nach dem Satz, der den Zielbegriff enthält, gesucht werden sollte.

3.3 Analyse der stilistischen Ähnlichkeit

Eine in der digitalen Philologie weit verbreitete Analysemethode vergleicht die Häufigkeiten von sehr vielen Einzelwörtern in mehreren Texten und ermittelt auf dieser Grundlage ein Maß für die (stilistische) Ähnlichkeit der Texte zueinander (die Methode wird meist Stilometrie genannt und verwendet unter anderem sogenannte Distanzmaße). Ein einschlägiger Anwendungsfall ist die Autorschaftsattribution, bei der Texte unbekannter oder umstrittener Autorschaft einem bekannten Autor zugeordnet werden sollen.[27]

Für solche Analysen ist es wichtig, nicht den gesamten in einer TEI-Datei enthaltenen Text zu berücksichtigen, sondern nur den Text, der auch tatsächlich vom mutmaßlichen Autor des fraglichen Textes geschrieben wurde und zum Haupttext gehört. Das bedeutet, man möchte Vorworte und Nachworte, aber sofern vorhanden auch editorische Anmerkungen von der Analyse ausschließen. (Technisch ausgedrückt: Man würde nur den Textinhalt des TEI `<body>`-Elements auswählen, dabei aber den Textinhalt aller vorhandenen `<note>`-Elemente löschen.) Denkbar wäre auch, zudem nur die Erzählerrede zu berücksichtigen, d.h. Überschriften und (direkte) Figurenrede auszuschließen, falls diese Information vorhanden ist. Bei Theaterstücken könnte man Bühnenanweisungen, Sprechernamen sowie Szenen- oder Akt-Angaben löschen. Vorausgesetzt, das TEI-kodierte Dokument enthält diese Informationen, ist ein solcher gezielter Zugriff auf den Text mit relativ einfachen Mitteln möglich.[28] Wollte man den gleichen Effekt mit einem unstrukturierten Textformat erreichen, müssten man mehrere separate Textfassungen erstellen und pflegen und jeweils die geeignete Textfassung auswählen.

[27] Einführend zu Stilometrie und Autorschaftsattribution: Hugh Craig, „Stylistic Analysis and Authorship Studies“, in *A Companion to Digital Humanities*, hrsg. von Susan Schreibman, Ray Siemens und John Unsworth (Oxford: Blackwell, 2004), 273–88; einen lesenswerten Überblick bietet Efstathios Stamatatos, „A Survey of Modern Authorship Attribution Methods“, *Journal of the Association for Information Science and Technology* 60, Nr. 3 (2009): 538–56, http://doi.org/10.1002/asi.v60:3. Für eine ausführliche Studie zu quantitativen Analysen auf stilistischer und thematischer Ebene, siehe Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Champaign: University of Illinois Press, 2013).

[28] Beispielsweise mit einem Python-Skript von wenigen Zeilen, dass das Modul „lxml“ nutzt.

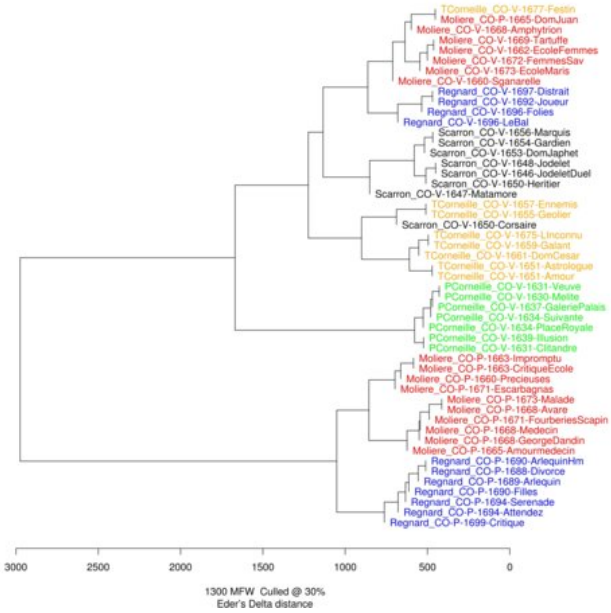

Abb. 5: Stilistische Ähnlichkeit von 54 französischen Komödien (Clustering in *stylo*)

Abbildung 5 zeigt ein Beispiel für eine solche Analyse, bei dem eine kleine Sammlung von französischen Theaterstücken auf ihre stilistische Ähnlichkeit hin untersucht wurde und wie eben beschrieben nur der Sprechertext für die Analyse berücksichtigt wurde. Es handelt sich um eine Reihe von Komödien verschiedener Autoren, die aufgrund der Häufigkeiten der 1100 häufigsten Wörter miteinander verglichen wurden. Je ähnlicher sich zwei Stücke stilistisch sind, desto näher stehen sie sich, vereinfacht gesagt, im Baumdiagramm. Auffallend ist, dass die wichtigste Dimension, nach der sich die Sammlung in zwei Gruppen gliedert, die Unterscheidung von Prosa und Vers ist und erst innerhalb dieser beiden Gruppen Autorschaft zum Tragen kommt (‚prose' und ‚vers' in den Beschriftungen der Stücke).[29]

[29] Das verwendete Tool ist stylo für R, siehe: https://sites.google.com/site/computationalstylistics/home. Die Parameter waren 1100 häufigste Wörter, 10 Prozent Culling, Distanzmaß Eder's Delta. Die zugrunde liegende Methode sowie das Beispiel werden ausführlicher diskutiert in: Christof Schöch, „Corneille, Molière et les Autres: stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik", in *Literaturwissenschaft im digitalen Medienwandel*, hrsg. von Christof Schöch und Lars Schneider, Beihefte von Philologie im Netz 7 (2014): 130–57,

3.4 Inhaltliche Analyse mit „Topic Modeling“

Für eine eine weitere, statistisch avancierte, auf den Textinhalt bezogene Analysemethode größerer Textsammlungen können Strukturinformationen in TEI-Dokumenten genutzt werden. Solche Analysen werden seit einigen Jahren vornehmlich nicht mehr über eine Stichwortsuche unternommen, da hier beispielsweise nur nach Einzelwörtern, nicht aber nach Themen gesucht werden kann. Auch selbst erstellte Listen von thematisch verwandten Wörtern sind vor dem Vorwurf der Voreingenommenheit nicht sicher. Stattdessen nutzt man bei dem hier gemeinten Verfahren des Topic Modeling die Information darüber, welche Wörter immer wieder in verschiedenen Textabschnitten gemeinsam auftreten, um Gruppen semantisch verwandter Wörter zu entdecken und ihre Verteilung in einer Textsammlung zu ermitteln. Die dafür eingesetzte Methode wird seit gut zehn Jahren intensiv entwickelt und genutzt.[30]

Um die Methode, die für kürzere Dokumente wie Zeitungsartikel oder wissenschaftliche Artikel entwickelt wurde, auch für sehr umfangreiche literarische Texte wie beispielsweise Romane einsetzen zu können, müssen solche Texte vorab in mehrere kürzere Segmente zerlegt werden. Ein denkbares Vorgehen dafür wäre, die Texte einfach in Segmente gleicher Länge zu zerteilen und die Länge der Segmente dabei als eine bestimmte Anzahl von Wörtern festzulegen, beispielsweise 2000 Wörter. Bei diesem Vorgehen werden allerdings vorhandene strukturelle Einheiten wie Kapitel oder Absätze ignoriert. Da man annehmen kann, dass Kapitel in Romanen oder Szenen und Akte in Theaterstücken eine gewisse thematische Einheit darstellen, erscheint es sinnvoll, diese Einheiten bei der Segmentierung zu berücksichtigen. Zumindest aber sollten Einheiten wie Absätze oder Figurenreden bei der Segmentierung nicht aufgespalten werden. Da diese Einheiten in TEI-kodierten Dokumenten in der Regel markiert sind, ist dies besonders einfach möglich. Es steht zu erwarten, dass die resultierenden „topics“ (etwa: Themen und Motive) dann eine größere Kohärenz oder zumindest eine größere Adäquatheit zu den zugrunde liegenden Texten haben werden, als bei

http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf.

[30] Zwei lesenswerte Einführungen in die Methode des Topic Modeling: David M. Blei, „Probabilistic Topic Models“, *Communications of the ACM* 55, Nr. 4 (2012): 77–84, http://doi.org/10.1145/2133806.2133826 sowie Mark Steyvers und Tom Griffiths, „Probabilistic Topic Models“, in *Latent Semantic Analysis: A Road to Meaning*, hrsg. von T. Landauer u.a. (Laurence Erlbaum, 2006).

willkürlicher Segmentierung. Letzteres ist allerdings bisher nicht systematisch untersucht worden.

Abb. 6: Visualisierung von Topics als Wordcloud

Abbildung 6 zeigt eine Visualisierung mehrerer Topics als sogenannte Wordcloud. Jeder Topic wird aus automatisch ermittelten, immer wieder gemeinsam auftretenden Wörtern gebildet, die verschiedene Arten von semantischer Kohärenz aufweisen können (abstrakte Themen, erzählerische Motive, Orte der Handlung, fremdsprachige Begriffe, etc.). Je größer ein Wort in der Wordcloud dargestellt ist, desto wichtiger ist es in dem entsprechenden Topic. Hier sind vier Topics dargestellt, die man stark vereinfachend mit den folgenden Begriffen zusammenfassen könnte (von links oben im Uhrzeigersinn): Tod, Meer, Schreiben, Heirat.[31]

3.5 Konfigurationsanalyse bei Theaterstücken

Seit Anne Ubersfeld in den 1970er Jahren das Aktanten-Modell für die Analyse von Theaterstücken vorgeschlagen bzw. angepasst und populär gemacht hat, hat die Literaturwissenschaft das Interesse an den Beziehungen zwischen den Figuren in einem Theaterstück und an den resultierenden Figurenkonstellationen nicht mehr verloren.[32] Dieses Interesse hat eine neue

[31] Dieses und weitere Beispiele werden ausführlicher kommentiert in: Christof Schöch, „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama“, *Digital Humanities Quarterly* (2015), http://digitalhumanities.org/dhq.

[32] Zum Thema Figurenkonstellationen, siehe: Anne Ubersfeld, *Lire le Théâtre*, nouv. éd. rev, Lettres Belin Sup (Paris: Belin, 1996); Solomon Marcus, *Mathematische Poetik*, übers. von Edith Mândroiu (Bucureşti; Frankfurt/Main: Editura Academiei; Athenäum Verlag, 1973); Manfred Pfister, *Das Drama: Theorie und Analyse* (München: W. Fink, 1977).

Qualität angenommen, seit Theaterstücke in zunehmender Anzahl digital und, im besten Falle, auch im TEI-Format vorliegen. Denn wenn explizit kodiert ist, welche Figuren in welcher Szene auf der Bühne sind und wieviel Sprechertext jeweils auf sie entfällt, wird eine ganz neue Form der Analyse von Figurenkonstellationen möglich.

Beispielsweise kann die Information darüber, welche Figuren sich wann und wie oft im Verlauf des Stückes treffen, automatisch erhoben werden und für eine graphische Repräsentation der Figurenkonstellation genutzt werden. Auch Informationen zur Dichte des Figurennetzwerkes und Vergleiche über Autoren oder Epochen hinweg, werden möglich. Die Beziehungen zwischen den Personen können darüber hinaus weiter im Sinne einer Intensität der Interaktion qualifiziert werden, weil sich berechnen lässt, wer in Anwesenheit welcher anderer Personen wie viel spricht. (Klar ist, dass ein solcher rein quantitativer Ansatz nichts darüber aussagt, wie mehr oder weniger wichtig oder entscheidend diese Interaktionen sind, noch zwangsläufig etwas darüber, wer über die reine gemeinsame Anwesenheit auch tatsächlich aktiv interagiert.)[33]

Abbildung 7 zeigt das Figurennetzwerk von Lessings *Nathan der Weise* (1779), wie es sich bei automatischer Extraktion der Anwesenheit der Figuren pro Szene mit geringer Nachbearbeitung ergibt.[34] Je häufiger sich zwei Figuren treffen, desto näher stehen sie sich im Netzwerk; je dicker die Linie zwischen zwei Figuren ist, desto mehr sprechen die Figuren miteinander.

Als Fazit für den Nutzen TEI-kodierter Texte für die Textanalyse gilt: Texte in TEI und nicht als plain text oder Word-Datei vorliegen zu haben, ermöglicht den Literaturwissenschaften den intelligenten Zugriff auf die „Textdaten", der eigentlich wünschenswert ist, sowohl was die Binnenstruktur der Texte angeht, als auch was Phänomene wie Namen, Orte, Zeiten und vieles mehr sowie Textvarianten angeht. Wenn dazu dann noch linguistische Annotationen kommen, in der TEI-Datei selbst (die das unterstützt) oder als se-

[33] Die Methode literaturwissenschaftliche Netzwerkanalyse wird ausführlich besprochen von: Peer Trilcke, „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft", in *Empirie in der Literaturwissenschaft*, hrsg. von Philip Ajouri, Katja Mellmann und Christoph Rauen (Münster: Mentis, 2013), 201–47.

[34] Die Abbildung stammt aus dem LINA-Projekt von Frank Fischer, Mathias Göbel, Dario Kampkaspar, Peer Trilcke (2015), siehe http://dlina.github.io/369. Bildlizenz: Creative Commons Attribution 4.0 International. Siehe auch: Peer Trilcke, Frank Fischer und Dario Kampkaspar, „Digitale Netzwerkanalyse Dramatischer Texte", in *DHd-Tagung* (Graz, 2015), http://gams.uni-graz.at/o:dhd2015.v.040.

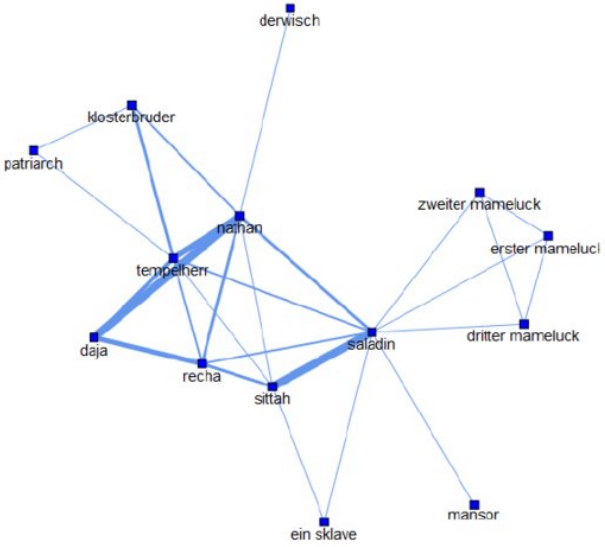

Abb. 7: Das Figurennetzwerk von Lessings *Nathan der Weise*

parate Annotation oder in einer anderen Form (beispielsweise in TCF), wird der Zugriff auf den Text noch differenzierter.

Fazit

Der vorliegende Beitrag hatte sich zum Ziel gesetzt, über adäquate digitale Repräsentationsformen von (literarischen) Texten für literaturwissenschaftliches Arbeiten zu informieren und einige der Möglichkeiten, die die Richtlinien der Text Encoding Initiative für die Textkonstitution und die Textanalyse anbieten, aufzuzeigen. Vor allem aber möchte der Beitrag damit ein Plädoyer leisten für die Nutzung von nicht-proprietären, semantisch orientierten, semi-strukturierten Textrepräsentationen wie TEI. Diese könnte man als „smart data" bezeichnen, weil sie nicht nur eine mechanische Reproduktion der Zeichenkette darstellen, sondern zahlreiche Annotationen zu bestimmten Textpassagen, Informationen über die Struktur des Textes sowie eine Kontextualisierung des Textes durch die Metadaten erlauben. Solche Daten bieten gegenüber anderen Repräsentationsformen von Text, wie digitale Faksimiles oder unstrukturierte, einfache Textdateien, erhebliche Vorteile. Die literaturwissenschaftliche Analyse kann umso nuancierter

und präziser sein, je mehr Informationen bei der Textkonstitution festgehalten wurden. Dieser arbeitsintensive Prozess muss im Übrigen nicht immer vollständig von Hand erfolgen, denn durch quantitative Verfahren identifizierte Phänomene (wie beispielsweise bei der automatischen Erkennung von Personennamen) können auch automatisch in TEI-kodierte Dokumente eingetragen, dort ergänzt oder korrigiert und für die Visualisierung oder weitere Analysen genutzt werden. Textkonstitution und quantitative Textanalyse können hier Hand in Hand gehen.[35]

Aus der Darstellung folgt einerseits, dass sich die literaturwissenschaftliche Forschung nicht mit Repräsentationsformen von Text zufrieden geben sollte, die nur wenig differenzierte Analysemöglichkeiten ermöglichen. Andererseits sollte das digitale Medium als literaturwissenschaftliches Forschungsinstrument aber auch nicht auf der Grundlage von bestimmten Repräsentationsformen von Text, die in der Tat inadäquat sind, im Ganzen verdammt werden. Vielmehr ist es notwendig, hier ein klares Desiderat für zukünftige Forschung in den Literaturwissenschaften zu erkennen. In der Tat ist die Verfügbarkeit wesentlicher Teile einer literaturhistorischen Texttradition in zuverlässigen, standardisierten, aktuellen wissenschaftlichen Ansprüchen genügenden digitalen Textfassungen derzeit ein Desiderat. Trotz vieler guter Initiativen liegen selbst für die meisten Autoren, deren Werke unstrittig zum engeren Kanon der Literaturgeschichtsschreibung gehören, solche digitalen Referenzausgaben nicht vor. Dies gilt in der Mehrzahl der literaturwissenschaftlichen Teilfächer einschließlich der Romanistik, allerdings mit Ausnahme der Klassischen Philologie und der Germanistik. Die Verfügbarkeit solcher digitalen Textfassungen wäre aber eine wesentliche Bedingung dafür, dass literaturwissenschaftlich interessante Fragestellungen auch in größeren Textmengen auf ebenso subtile wie philologisch verlässliche Weise bearbeitet werden können. Zudem gilt: Je mehr Texte in TEI-kodierter Form vorhanden sind, desto größer werden die Vorteile der technischen Kompatibilität sowie der inhaltlichen Vergleichbarkeit der Texte für Analyse, Interpretation und Literaturgeschichtsschreibung. Denn wenn digitale Texte nicht in vielen unterschiedlichen Forma-

[35] Zum Begriff der „Daten" in den Geisteswissenschaften, siehe Trevor Owens, „Defining Data for Humanists: Text, Artifact, Information or Evidence?" *Journal of Digital Humanities* 1, Nr. 1 (2011), http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens sowie Christof Schöch, „Big? Smart? Clean? Messy? Data in the Humanities", *Journal of the Digital Humanities* 2, Nr. 3 (2013): 2–13, http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities.

ten, sondern in einem Standard wie TEI kodiert vorliegen, können Texte aus verschiedenen Quellen leichter zusammengeführt und so für spezifische Fragestellungen geeignete Sammlungen leichter erstellt werden. Das bedeutet, dass TEI nicht nur verstärkt genutzt werden sollte, sondern mehr noch, dass es der selbstverständliche Standard für die Repräsentation von Texten sein sollte. Damit hängt zusammen, dass verstärkt Kompetenzen im Bereich der digitalen Textkodierung und der Nutzung der Text Encoding Initiative als grundlegendes literaturwissenschaftliches Arbeitsinstrument im Studium und in der Doktorandenausbildung vermittelt werden sollten. Erst dann kann die Vision einer digitalen Literaturwissenschaft auf Grundlage umfangreicher Sammlungen verlässlicher Texte Realität werden.

*
**

Anhang: Hinweise für die weitere Beschäftigung mit TEI

Digitale Textedition

Burnard, Lou. *What is the Text Encoding Initiative?* Marseille: OpenEdition, 2014, http://books.openedition.org/oep/679.

———, Katherine O'Brien O'Keeffe und John Unsworth, Hrsg. *Electronic Textual Editing*. New York: MLA, 2006.

Hockey, Susan. *Electronic Texts in the Humanities*. Oxford: Oxford Univ. Press, 2000.

Sahle, Patrick. „Digitale Editionstechniken". In: *Digitale Arbeitstechniken für die Geistes- und Kulturwissenschaften*, hrsg. von Martin Gasteiner und Peter Haber, 231–49. Wien: UTB, 2009.

Shillingsburg, Peter. *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge: Cambridge Univ. Press, 2006.

Quantitative Textanalyse

Adolphs, Svenja. *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*. London und New York: Routledge, 2006.

Blei, David M. „Probabilistic Topic Models". *Communications of the ACM* 55, Nr. 4 (2012): 77–84.

Jannidis, Fotis. „Methoden der computergestützten Textanalyse". In *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*, hrsg. von Ansgar und Vera Nünning, 109–32. Stuttgart: Metzler, 2010.

Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History*. Champaign: Univ. of Illinois Press, 2013.

Stamatatos, Efstathios. „A Survey of Modern Authorship Attribution Methods". *Journal of the Association for Information Science and Technology* 60, Nr. 3 (2009): 538–56.

Trilcke, Peer. „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft". In *Empirie in der Literaturwissenschaft*, hg. von Philip Ajouri, Katja Mellmann und Christoph Rauen. (Münster: Mentis, 2013): 201–47.

Beispiele für digitale Texteditionen in TEI

Ainsworth, Peter und Godfried Croenen, Hrsg. *The Online Froissart: A Digital Edition of the Chronicles of Jean Froissart*. Univ. Sheffield, www.hrionline.ac.uk/onlinefroissart/index.jsp.

Bohnenkamp-Renken, Anne, Silke Henke und Fotis Jannidis, Hrsg. *Goethes Faust: eine genetische Edition*. Würzburg, 2015, http://faustedition.uni-wuerzburg.de.

Emsley, Clive, Tim Hitchcock und Robert Shoemaker, Hrsg. *The Proceedings of the Old Bailey*. 2003–2015, www.oldbaileyonline.org.

Ertler, Klaus-Dieter, Alexandra Fuchs, Michaela Fischer und Elisabeth Hobisch, Hrsg. *Moralische Wochenschriften*. Univ. Graz, 2011, http://gams.uni-graz.at/context:mws.

Jansen, Leo, Hans Luijten und Nienke Bakker, Hrsg. *Vincent Van Gogh, The Letters*. Van Gogh Museum / Huyghens ING, 2014, http://vangoghletters.org/vg.

Rosselli Del Turco, Roberto, Hrsg. *The Digital Vercelli Book*. Pisa: CISIAU/Laboratorio di Cultura Digitale, 2013, http://vbd.humnet.unipi.it.

Tuck, John und Ronald Milne et al. *Codex Sinaiticus*. London: British Library, 2007. http://codexsinaiticus.org.

Vauthier, Bénédicte, Hrsg. *Manuscrito digital de Juan Goytisolo*, Univ. Bern, 2013. http://goytisolo.unibe.ch.

Wiering, Frans, Hrsg.*Thesaurus musicarum italicarum online*, Univ. Utrecht, 2000–2005. http://tmiweb.science.uu.nl.

Siehe auch: Patrick Sahle, *A Catalogue of Digital Scholarly Editions*, 2008–2015, www.digitale-edition.de/index.html.

Quellen für größere Textsammlungen in TEI

Digitale Bibliothek, TextGrid, www.textgridrep.de.

Deutsches Textarchiv, BBAW, www.deutschestextarchiv.de.

Théâtre classique, hrsg. v. Paul Fièvre, Paris-Sorbonne, www.theatre-classique.fr.

Biblioteca italiana, Università di Roma-Sapienza, www.bibliotecaitaliana.it.

Oxford Text Archive, Oxford University, http://ota.ox.ac.uk.

Weitere Ressourcen: Tools und Handreichungen

TEI by Example, hrsg. von Melissa Terras und Edward Vanhoutte, http://teibyexample.org (interaktives Tutorial).

Digitale Textedition mit TEI, Redaktion Christof Schöch, Göttingen: DARIAH-DE, 2014, https://de.dariah.eu/tei-tutorial (Schulungsmaterialien).

TXM: http://textometrie.ens-lyon.fr (Analysetool, das XML/TEI nutzt).

TXM-Kurzreferenz, von Christof Schöch, 2014 https://zenodo.org/record/10769 (Handreichung für den Einstieg in TXM).

TAPAS (TEI Archive, Publishing, and Access Service), Brown Univ., www.tapasproject.org.

jEdit mit XML-Plugins, http://jedit.org (kostenloser Editor, der XML/TEI unterstützt).

oXygen Editor, www.oxygenxml.com (sehr leistungsfähiger Editor mit spezifischer Unterstützung von TEI).

TextGrid Lab, https://textgrid.de (umfassende Arbeitsumgebung für digitale Editionen).