BMBF eHumanities Nachwuchsforschergruppe: „Computergestützte literarische Gattungsstilistik“

Lehrstuhl für Computerphilologie
Universität Würzburg
http://www.clgs.hypotheses.org

Einleitung

Die weltweit vorangetriebene Digitalisierung des kulturellen Erbes hat im Bereich der Volltextdigitalisierung inzwischen einen Umfang erreicht, der neue methodische Zugänge zu literaturwissenschaftlichen Fragestellungen ermöglicht und erfordert, woraus zugleich neue Fragestellungen entstehen. Übergeordnete Zielsetzung der Nachwuchsforschergruppe „Computergestützte literarische Gattungsstilistik“ ist es, eine methodische Konvergenz herzustellen zwischen neuesten Verfahren der quantitativen Analyse literarischer Texte und grundlegenden literaturwissenschaftlichen Fragestellungen aus dem Bereich der Gattungstheorie und der Stilistik.

Eine solche Konvergenz herzustellen, bedeutet konkret, dass zwei Arten von Wissen in Beziehung zueinander gesetzt werden müssen: Wissen über historische oder fachwissenschaftliche Kategorisierungen (Theorie und Geschichte der literarischen Gattungen und stilistischer Phänomene) einerseits, und Wissen über die Kategorien und Differenzierungen, die sich aus der quantitativen Analyse stilistischer Merkmale (unterschiedlicher Ebene, Komplexität und Kombination) ergeben, andererseits.

Die Nachwuchsgruppe wird zwischen romanistischer Literaturwissenschaft und angewandter Informatik angesiedelt sein: sie geht von etablierten literaturwissenschaftlichen Fragestellungen grundlegender Bedeutung aus, allerdings von Anfang an in der Perspektive einer computergestützten Methodik. Ziel ist es, die literaturwissenschaftlichen Fragestellungen durch eine Kombination umfangreicher Textdaten, innovativer Analysemethoden und hermeneutischer Kontextsensibilität auf neuer Grundlage und mit einem neuen Blick beantworten zu können. Dies wird anhand mehrerer umfangreicher, digital vorliegender Textsammlungen aus dem Bereich des französischen Theaters des 17.-18. Jahrhunderts sowie des französischen und spanischen Romans des 18.-19. Jahrhunderts unternommen.

1. Wissenschaftliche Arbeitsziele

Die Nachwuchsgruppe wird an aktuelle Forschung zu Gattungstheorie und Stilistik anknüpfen und neueste Verfahren der computergestützten Textanalyse anwenden und weiterentwickeln. Spezifische Fragestellungen werden unter Verwendung bestimmter Verfahren an eine Reihe verfügbarer Textsammlungen herangetragen. Einzelne Teilprojekte beziehen sich auf unterschiedliche Teile der Textsammlungen (u.a. Untergattungen des Romans und des Dramas), unterschiedliche Epochen (im Zeitraum 17.-19. Jahrhundert) und Nationalliteraturen (Frankreich und Spanien, sowie vergleichend auch Deutschland).

Die methodischen und fachwissenschaftlichen Ziele der Gruppe betreffen zunächst die Klärung der literaturtheoretischen Frage, wie die Beziehung von Stil und Gattung sinnvoll zu konzeptualisieren ist und auf welcher stilistischen Grundlage Gattungsunterscheidungen getroffen werden können. Was sind Gattungen und welche Attribute oder „Facetten“ ((Kessler et al. 1998) machen verschiedene Gattungen aus? Und welche automatisch identifizierbaren, stilistischen Eigenschaften oder Indikatoren, auf welchen sprachlichen Ebenen, sind jeweils Indikatoren für diese Attribute? Oder generischer gefasst, wie kann eine Verifikations- bzw. Falsifikationskette von der Theorie (bspw. Konzept des *roman libertin*) über bestimmte

Hypothesen (konkrete Merkmale des *roman libertin*) zu spezifischen Indikatoren (stilistische Einzelmerkmale) gespannt werden? (vgl. Jannidis 2010). Welche Beziehungen bestehen zwischen einzelnen Gattungen und Untergattungen, welche Überschneidungen gibt es zwischen verwandten Untergattungen? Wie hängen Gattungsstil und Epochen-/Zeitstil (vgl. Müller 2009) zusammen, welche historischen Determinanten für Gattungen (vgl. Wolf 2009) sind zu beobachten? Hier gilt es auch, methodische Lösungen zur Trennung von konkurrierenden Signalen zu entwickeln (bspw. Autoren- vs. Gattungssignal: vgl. Kestemont 2012, Schöch 2013a). Die Identifikation geeigneter Merkmale kann iterativ erfolgen, indem zunächst explorativ Klassifikations-Methoden angewandt und dann die für eine bestimmte Klassifikation entscheidenden Merkmale identifiziert werden („feature selection"). Lassen sich unter diesen Merkmalen Gemeinsamkeiten oder Muster identifizieren, können über „pattern generalization" (Lin 1998) neue Merkmalsbündel generiert und für eine erneute Klassifikation genutzt werden.

Die Ausgangshypothese der Teilprojekte wird in der Regel eine begründete Annahme über einen Bezug zwischen einer Reihe von gattungsbezogenen Einzelindikatoren und bestimmten gattungsbezogenen Facetten für mehrere Gattungen bzw. Untergattungen sein, die dann unter Anwendung quantitativer Methoden auf mehr oder weniger intensiv annotierte Textsammlungen formalisiert, modelliert, überprüft, modifiziert und interpretiert wird. Der Ansatz erfordert (a) die Formalisierung von Merkmalen für die computergestützte Analyse, (b) die Nachvollziehbarkeit der statistischen Verfahren, und (c) die Interpretierbarkeit der Merkmalsbündel in Bezug auf Funktionen und Kontexte der in Frage stehenden Gattungen.

Bezüglich der eingesetzten Methoden und Verfahren kommen der Natur der Fragestellungen entsprechend (die letztlich Klassifikationsprobleme sind) überwachte Klassifikations- und nicht-überwachte Clusteringmethoden aus dem Bereich des Text Mining und Machine Learning in Frage. Machine Learning (vgl. Han et al. 2011, Witten et al. 2011) ist ein flexibler, generischer Ansatz zum Umgang mit hochdimensionalen Daten. Zwei zentrale Verfahren sind stilometrische Cluster-Analyse und Topic Modeling. Die stilometrische Cluster-Analyse (Juola 2006) ist ein Verfahren des nicht-überwachten Clusterings umfangreicher Textsammlungen aufgrund von Merkmalen wie Wort- oder Zeichenhäufigkeit. Topic Modeling (Blei 2011) ist ein Verfahren, das es erlaubt, inhaltliche Muster in kleineren oder größeren Textabschnitten zu identifizieren.

Alle Verfahren müssen für die analysierten Gattungen und Sprachen optimiert werden. Dies ist neben der Frage der Wahl der zu berücksichtigenden sprachlichen Merkmale und Parameter vor allem durch Auswahl und Optimierung von Algorithmen möglich, was auch die Validierung der Verfahren einschließt. Komplexere Verfahren der computergestützten Stilistik erfordern für optimale Ergebnisse die semi-automatische Auszeichnung eines Lernkorpus nach einem breiten Spektrum linguistischer Merkmale. Zudem spielen Visualisierungsverfahren eine wichtige Rolle: sie werden nicht nachträglich llustrativ, sondern iterativ und heuristisch eingesetzt, denn sie erlauben häufig die Entdeckung von Mustern, Trends und Besonderheiten in den Ergebnissen. Aus der Verknüpfung von Metadaten, linguistischer Annotation, verschiedenen Analyse-verfahren und Visualisierungstechniken enstehen strategisch einsetzbare, komplexen Verfahren.

Mehrere Sammlungen digitalisierter Volltexte werden der Gruppe als gemeinsame Grundlage für die einzelnen Forschungsvorhaben zur Verfügung stehen; diese werden großenteils während der Vorphase aufbereitet und verfügbar gemacht:

- Französisches Drama, etwa 1610-1795. Umfang etwa 600 Einzeltexte, verfügbar unter www.theatre-classique.fr.
- Spanisches Drama, etwa 1600-1800. Umfang etwa 200 Einzeltexte aus verschiedenen Quellen.
- Französischer Roman, etwa 1780-1920. Umfang etwa 600 Einzeltexte aus verschiedenen Quellen.
- Spanischer Roman, etwa 1800-1920. Umfang etwa 200 Einzeltexte aus verschiedenen Quellen.

Diese Sammlungen sind ausreichend umfangreich, dass auch Teile der Sammlungen analysiert und Verlaufsstudien vorgenommen werden können. Es wird nur auf bereits im Volltext verfügbare Texte zurückgegriffen, die allerdings teilweise gesammelt und in ein geeignetes, zentrales Format überführt werden müssen. Je nach den Interessen und dem disziplinären Hintergrund der Beteiligten können die Textsammlungen in der Hauptphase erweitert und durch weitere Sammlungen ergänzt werden. Hierzu gehört auch, eine nachhaltige Repositoriums-Struktur aufzubauen. Hier werden die Texte (mit Metadaten und linguistischer Annotation) in einem zentralen Datenformat archiviert, aus dem jeweils für bestimmte Analyseverfahren geeignete Formate und Sammlungen dynamisch generiert werden.

2. Stand der Forschung

Die für die Bearbeitung der Forschungsfragen in Frage kommenden Methoden aus dem Bereich der Digital Humanities (DH) gehören zum Bereich Text Classification, Text Mining und Machine Learning (hierbei insbesondere diverse Verfahren der Dimensionalitätsreduktion (Distanzmaße, Clustering, Principal Component Analysis) sowie Regressionsanalysen sowie Topic Modeling. Die Gruppe kann im Kontext dieser Methoden und ihrer Anwendung auf geisteswissenschaftliche Fragen an mehrere Forschungsstränge im Bereich der DH und der Informatik anknüpfen. Erstens an Forschung zu Theorie, Anwendungsfällen und methodischen Fragen quantitativer Verfahren der literarischen Textanalyse (vgl. Adolphs 2006, Brunet 2011, Ramsay 2011), speziell der Klassifikation von Texten auf der Grundlage stilistischer Gemeinsamkeiten (Craig et al. 2009; Rybicki & Eder 2011; Jockers 2013).

Eine sich aktuell verstärkende Tendenz ist die Berücksichtigung, über Fragen der Autor-Attribution hinaus, von weiteren Faktoren wie Gattung, Epoche oder Geschlecht: in Bezug auf Gattungen mit literaturwissenschaftlicher Perspektive behandeln bisherige neuere Arbeiten unter anderem die Frage der diachronen Entwicklung von Gattungen (Moretti 2005) und den Einsatz von Cluster Analyse für die Gattungsklassifikation (Allison et al. 2011) oder spezifischer die Erprobung der "unmasking"-Prozedur (Kestemont et al. 2012) und die Abgrenzung einer spezifischen Untergattung des Romans von Romanen der Höhenkamm-Literatur auf Grundlage syntaktischer Komplexitätsmaße (Jautze et al. 2013).

Im Bereich der Corpuslinguistik geht die computergestützte Untersuchung der stilistischen Unterschiede von (literarischen) Gattungen und Untergattungen bis in die 1980er-Jahre zurück, mit Pionierarbeiten von Douglas Biber zur Modellierung des Zusammenhangs zwischen funktionalen Gattungsaspekten und stilistischen Merkmalen, die zu synthetischen Dimensionen zusammengefasst werden (Biber 1989; Biber 1992) und der Erprobung einer breiten Auswahl von potentiellen "style markers" (Karlgren & Cutting 1994). Außerdem wurden bspw. die vergleichende Evaluation von token-basierten, syntaktischen und anderen Merkmalen vorgenommen (Wolters & Kirsten 1999, Stamatatos et al. 2000) und verschiedene Klassifikations-Algorithmen erprobt (Snyman et al. 2011). Darüber hinaus wird die Gruppe allgemein an Forschung aus dem Bereich der Korpuslinguistik und der *corpus stylistics* (u.a. Leech 2008, Fischer-Starke 2010, Biber 2011; einführend: Biber & Conrad 2009) anschließen können.

Ein weiterer Anknüpfungspunkt ist stärker informatisch und statistisch fokussierte Forschung, die Verfahren der quantitativen Textanalyse oder spezifische Algorithmen evaluiert (Yang 1999; Jockers & Witten 2010; vgl. Juola 2006) oder neue Tools entwickelt sowie informatiknahe Forschung zu Verfahren und Einsatz der Visualisierung für die Interpretation von Zwischenergebnissen (vgl. Tufte 2001, Puretskiy et al. 2010). Zwei aktuelle Desiderate, die die Gruppe einlösen wird, betreffen die Nutzung semantisch orientierter Verfahren wie Topic Modeling für eine quantitative Stilistik auf der Ebene der Lexik und Semantik sowie den Mangel an Erfahrungen mit Sprachen wie Französisch oder Spanisch, aber auch Deutsch, d.h. die Übertragung und Anpassung von Verfahren, für die Erfahrungswerte mit englischen Texten vorliegen, auf die in der Gruppe behandelten Sprachen.

Die Gruppe wird auf den Stand der Forschung im Bereich der literarischen Stilistik und der Gattungstheorie aufbauen. Bezugspunkt ist weniger der individuelle Werk- oder Autorenstil oder die Stilinterpretation (in der Tradition von Léo Spitzer oder Pierre Guiraud; vgl. Karabétian 2000), vielmehr geht es um einen induktiven, deskriptiven Blick auf die stilistischen Merkmale literarischer Gattungen und Untergattungen sowie deren historische Entwicklung. Eine wichtige neuere Entwicklung betrifft hier die Auffassung von Stil als relationales Phänomen (Sandig 2006:85ff, Argamon & Koppel 2010): Stil konturiert sich zwar nicht als Abweichung von einer (problematischen) generellen Norm, wohl aber relativ zu anderen Texten oder Textmustern (wie einer Gattung; Beispiel: Leech 2008).

Hier kann die computergestützte Stilistik ansetzen, denn sie beruht auf der Festellung von Ähnlichkeiten und Differenzen zwischen Texten. Eine zweite Entwicklung betrifft die Auffassung von Stil als „Bündel kookkurrierender Merkmale“ (Sandig 2006:55), die in der Regel auf unterschiedlichen Beschreibungsebenen (Lexik, Morphologie, Syntax, Textkohäsion; quer dazu: Rhetorik; außerdem Diskurstypen) angesiedelt sind. Ein solcher offener Stilbegriff ist nicht nur geeignet, Stilistik und Gattungstheorie zusammenzubringen und damit die von Dominique Combe eingeforderte „stylistique des genres“ (Combe 2002; grundlegend auch Larthomas 1998) zu realisieren, sondern diese Gattungsstilistik auch computergestützt vorzunehmen. Das computergestützt erstmalig einlösbare Desiderat liegt darin, induktiv und umfassend zahlreiche Merkmale, sowohl in ihrer gegenseitigen Abhängigkeit als auch in ihrer jeweiligen Gewichtung, erfassen und bewerten zu können.

Auch die neuere Gattungstheorie ist hier anschlussfähig. Es hat sich die Auffassung durchgesetzt, dass literarische Gattungen sich nicht mit einem idealistischen, deduktiven Ansatz systematisieren lassen (Schaeffer 1989, Zymner 2003). Vielmehr sind sie als historische Konventionen zu verstehen, die komplexe und sich dynamisch entwickelnde „generic facets“ (Kessler et al. 1998) umfassen. Diese beziehen sich zu unterschiedlichen Anteilen auf Themen, Plot und diverse stilistische Merkmale (vgl. Hoffmann 2009). Die quantitative Stilistik verspricht, die Berücksichtigung feinster stilistischer Merkmale mit zeitlichen oder gattungsbezogenen Entwicklungen zu verbinden. Hier setzen auch Beiträge aus computerlinguistischer Perspektive (Pionier: Biber 1992; neuer: Argamon & Dodick 2004) und neueste quantitative Beiträge zur historischen Gattungsstilistik an (Jannidis & Lauer 2013, Jockers 2013).

3. Qualifikationskonzept

Neben der formalen Weiterqualifikation im Rahmen der Promotions- und Habilitationsverfahren wird die Nachwuchsgruppe ein Weiterbildungskonzept entwickeln, in dem die Gruppenmitglieder Kompetenzen aus den Bereichen Literaturwissenschaft und Informatik gemeinsam erarbeiten können. Dies betrifft unter anderem Grundbegriffe der Stilistik und Gattungstheorie; Kenntnisse in Statistik und Datenanalyse; Kenntnisse in XML-Technologien zur Strukturierung und Anreicherung von Texten; sowie Daten- und Projektmanagement. Diese Kompetenzen werden auch durch interne und externe Workshops aufgebaut, wofür in der Vorphase ein anwendungsorientiertes Konzept entwickelt wird.

Im Sinne einer Publikationsstrategie für die Gruppe wurden für unterschiedliche Teilaspekte des Projekts bzw. je nach literaturwissenschaftlichem oder informatischem Schwerpunkt geeignete Konferenzen und Zeitschriften identifiziert. In unterschiedlichen Konstellationen sollen hier jeweils gemeinsame Vorträge und kollaborative Veröffentlichungen entstehen. Die Mitglieder werden ihrer Qualifikationsphase entsprechend in die Networking-, Vortrags- und Publikationsaktivitäten der Gruppe eingebunden.

Im Sinne der Sichtbarkeit nach Außen werden auch digitale Publikations- und Kommunikationsmedien (Projektblog, öffentliche kollaborative Bibliographie) genutzt. Die Koordination und Dokumentation wird über das bewährte Mittel eines nichtöffentlichen Projektwikis umgesetzt.

4. Organisatorisches

In der derzeit geförderten Vorbereitungsphase, die von April 2014 bis März 2015 läuft, werden geeignete Personen als mögliche Mitglieder der Nachwuchsforschergruppe rekrutiert. Gemeinsam mit den designierten Mitgliedern und dem Mentor sowie assoziierten Personen werden die einzelnen Forschungsbereiche entwickelt, abgestimmt und geplant. Mit Unterstützung durch Hilfskräfte wird ein Teil der vorgesehenen Textsammlungen aufbereitet sowie ein geeignetes Text-Repositorium aufgebaut. Zudem werden die methodischen und literaturwissenschaftlichen Fragestellungen und Lösungsansätze weiter konkretisiert und soweit möglich exemplarisch erprobt und in der Fachcommunity vorgerstellt. Schließlich werden die Kontakte zu den Kooperationspartnern gepflegt und Vereinbarungen über die konkrete Form der Kooperation getroffen.

Die bereits beschriebenen literaturwissenschaftlichen und methodischen Fragestellungen der Gruppe werden in der Vorbereitungsphase so konkretisiert und in teilautonome Aufgaben gegliedert, dass jedes Mitglied im Rahmen des gemeinsamen Themas und Methodenrepertoires eine spezifische, auch fachwissenschaftlich eigenständige Fragestellung bearbeiten und spezifische methodische Lösungen entwickeln kann. Das Forschungsthema, das für die informatische Stelle vorgesehen ist, wird so definiert, dass sowohl eine aus der Perspektive der Forschung zu Machine Learning interessanter Problemkomplex bearbeitet werden, als auch Beiträge zum Rahmenthema der Gruppe und/oder zu Einzelprojekten entstehen können.

In der Hauptphase, deren Beantragung und Bewilligung noch aussteht und die für eine Laufzeit von vier Jahren ausgelegt ist, wird die Nachwuchsgruppe aus den folgenden Personen bestehen: einem Leiter der Nachwuchsgruppe (Literaturwissenschaft), einer Postdoktorandenstelle (Informatik), drei-vier Doktorandenstellen (Literaturwissenschaft, evtl. Informatik), sowie evtl. die Stelle eines technischer Mitarbeiter (Fachinformatik).

Die Nachwuchsgruppe wird an der Universität Würzburg angesiedelt und administrativ angebunden sein. Prof. Fotis Jannidis (Lehrstuhl für Computerphilologie) wird die Rolle des Mentors übernehmen. Ebenfalls beteiligt sind Prof. Dr. Brigitte Burrichter (Institut für Romanistik) und Prof. Dr. Andreas Hotho (Institut für Informatik). Die Qualifikationsverfahren können von der Philosophischen Fakultät betreut werden, bei entsprechenden Themen ergänzt durch die Fakultät für Mathematik und Informatik.

An der Universität Würzburg sind in idealer Weise die Voraussetzungen dafür gegeben, dass die Nachwuchsgruppe erfolgreich wird arbeiten können. Die Gruppe wird eingebettet sein in einen Lehr- und Forschungskontext, der unter anderem aktuelle Forschungen zur Textanalyse, laufende Infrastrukturprojekte wie TextGrid und DARIAH-DE, den interdisziplinären Studiengang „Digital Humanities“ (BA und MA) und ein Digital Humanities-Zentrum umfasst. Die Aktivitäten der Gruppe können auch in das Lehrangebot am Lehrstuhl für Computerphilologie und am Institut für Romanistik der Universität Würzburg einfließen. Die Nachwuchsgruppe wird auch über Würzburg hinaus in Deutschland und in Europa gut vernetzt sein und zahlreiche bestehende Kontakte nutzen und intensivieren können.

Projektleitung
Dr. Christof Schöch
Universität Würzburg
Institut für Deutsche Philologie
Lehrstuhl für Computerphilologie
Am Hubland - 97074 Würzburg
Telefon: +49-(0)931-31-85704
Email: christof.schoech@uni-wuerzburg.de
Web: http://www.clgs.hypotheses.org und http://kurzlink.de/gattungsstilistik

Bibliographie

Adam, Jean-Michel (2001). *Les textes: types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris: Armand Colin.

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

Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore (2011). *Quantitative Formalism: An Experiment*. Stanford: Standford Literary Lab.

Argamon, Shlomo & M. Koppel (2010). "The Rest of the Story: Finding Meaning in Stylistic Variation", in: *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*, hg. von Argamon, Burns & Dubnov. Berlin: Springer, 79-112.

Argamon, Shlomo, and Jeff Dodick (2006). „Conjunction and Modal Assessment in Genre Classification", in: *Computing Attitude and Affect in Text: Theory and Applications*, hg. von James G. Shanahan, Janyce Wiebe und Yan Qu. Heidelberg: Springer, 2006, 1-8.

Biber, Douglas (1992). "The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding", in: *Computers in the Humanities*, 26.5-6, 331-347.

Biber, Douglas (2011). "Corpus Linguistics and the Study of Literature: Back to the Future?" *Scientific Study of Literature* 1.1, 15-23. doi:10.1075/ssol.1.1.02bib.

Biber, Douglas, and Susan Conrad (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Blei, David M. (2011). „Introduction to Probabilistic Topic Models", *Communication of the ACM*.

Brunet, Etienne (2011). *Ce qui compte, écrits choisis. Tome II, méthodes statistiques*. Paris: Champion.

Brunner, Annelen (2013). "Automatic recognition of speech, thought, and writing representation in German narrative texts". *Literary and Linguistic Computing*, 28.4, 563-575.

Brunner, Annelen (2012). *Automatische Erkennung von Redewiedergabe in literarischen Texten*. Universität Würzburg (unpublizierte Dissertation).

Burges, Christopher. "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Mining and Knowledge Discovery* 2 (1998): 121–167.

Burrows, John (2002). "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17.3, 267-287. doi:10.1093/llc/17.3.267.

Craig, Hugh et al. (2009). *Shakespeare, Computers and the Mystery of Authorship*. Cambridge: Cambridge Univ. Press.

Craig, Hugh (2010). Intelligent Archive. http://www.newcastle.edu.au/school/hss/research/groups/cllc/intelligent-archive.html.

Combe, Dominique (2002). "La stylistique des genres", in: *Langue française* 135, 33-49.

Eder, Maciej & Jan Rybicki (2011). Computational Stylistics Scripts for R. https://sites.google.com/site/computationalstylistics/.

Fischer-Starcke, Bettina (2010). *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries*. Continuum, 2010.

Han, Jiawei, Micheline Kambe & Jian Pei (2011). *Data Mining: Concepts and Techniques*. 3rd ed. Burlington, MA: Elsevier, 2011.

Heiden, Serge, Jean-Philippe Magué, & Bénédicte Pincemin (2010). "TXM : Une Plateforme Logicielle Open-source Pour La Textométrie - Conception et Développement", in: *Statistical Analysis of Textual Data – Proceedings of JADT 2010*, 2:1021–1032. http://halshs.archives-ouvertes.fr/halshs-00549779.

Hoffmann, Michael (2009). "Mikro- und makrostilistische Einheiten im Überblick", in: *Rhetorik und Stilistik, Ein internationales Handbuch historischer und systematischer Forschung*, Band 2, hg. von Ulla Fix, Andreas Gardt & Joachim Knape. Band 2, Berlin: de Gruyter, 1529-45.

Jannidis, Fotis (2010). "Methoden der computergestützten Textanalyse", in: *Methoden der literatur- und kulturwiss. Textanalyse*, hg. von Vera Nünning & Ansgar Nünning. Stuttgart & Weimar: Metzler, 109-132.

Jannidis, Fotis & Gerhard Lauer (2013). "Burrows Delta and its Use in German Literary History". In: *Distant Readings - Descriptive Turns. Topologies of German Culture in the Long Nineteenth Century*, hg. von Matt Erlin & Lynne Tatlock. Rochester: Camden House (im Erscheinen).

Jannidis, Fotis & Christof Schöch (2013). "Report on the DARIAH-DE Expert Workshop Quantitative Text Analysis for Literary History", in: *DARIAH-DE Papers* (im Erscheinen).

Jautze, Kim, Corina Koolen, Andreas van Cranenburgh, & Hayco de Jong (2013). "From High Heels to Weed Attics: A Syntactic Investigation of Chick Lit and Literature." In *Proceedings of the Second Workshop on Computational Linguistics for Literature*, Atlanta, Georgia, 2013, 72-81.

Jockers, Matthew & Daniela Witten (2010). "A Comparative Study of Machine Learning Methods for Authorship Attribution." *Literary and Linguistic Computing* 25.2, 215-223.

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

Juola, Patrick (2006). “Authorship Attribution.” *Foundations and Trends in Information Retrieval* 1.3, 233-334.

Karabétian, Étienne (2000). *Histoire Des Stylistiques*. Paris: Armand Colin.

Karlgren, Jussi & Douglas Cutting (1994). “Recognizing text genres with simple metrics using discriminant analysis”. In *Proceedings of the 15th conference on Computational linguistics (COLING '94)*, Vol. 2, 1071-1075.

Kessler, Brett, Geoffrey Numberg, and Hinrich Schütze (1998). “Automatic Detection of Text Genre.” In *Proceedings of ACL 1998*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1997, 32-38. doi:10.3115/976909.979622.

Kestemont, Mike, Kim Luyckx, Walter Daelemans, and Thomas Crombez (2012). “Cross-Genre Authorship Verification Using Unmasking.” *English Studies* 93.3, 340–356. doi:10.1080/0013838X.2012.668793.

Larthomas, Pierre (1998). *Notions de stylistique générale*. Paris: PUF.

Leech, Geoffrey (2008). "Work in progress in corpus stylistics: a method of finding 'deviant' or 'key' features of texts, and its application to 'The Mark on the Wall'", in: Language in Literature. Style and foregrounding. London: Longman, 162-178.

Lin, Dekang (1998). “Automatic Retrieval and Clustering of Similar Words”, in: *COLING-ACL98*, Montreal, Canada.

Moretti, Franco (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Müller, Wolfgang (2009). „Epochenstil/Zeitstil“, in: *Rhetorik und Stilistik, Ein internationales Handbuch historischer und systematischer Forschung*, hg. von Ulla Fix, Andreas Gardt & Joachim Knape. Band 2, Berlin: de Gruyter, 1271-85.

Puretskiy, Andrey A., Gregory L. Shutt, & Michael W. Berry (2010). “Survey of Text Visualization Techniques.” In *Text Mining*, hg. von Michael W. Berry & Jacob Kogan. Wiley, 105-127.

Ramsay, Stephen (2011). *Reading Machines : Toward an Algorithmic Criticism*. Urbana Ill.: University of Illinois Press.

Rybicki, Jan & Maciej Eder (2011). “Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?” *Literary and Linguistic Computing* 26.3, 315-321.

Sandig, Barbara (2006). *Textstilistik des Deutschen*. 2. Auflage. Berlin: de Gruyter.

Schaeffer, Jean-Marie (1989). *Qu’est-ce qu’un genre littéraire?* Paris: Seuil, 1989.

Schöch, Christof (2011). *La Description double dans le roman français des Lumières, 1760-1800*. Paris: Classiques Garnier.

Schöch, Christof (2013a). “Fine-tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater”, *Digital Humanities Conference 2013*, http://dh2013.unl.edu/.

Schöch, Christof (2013b). “Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik”, in: *Revolution der Medien, Evolution der Literaturwissenschaft?*, hg. von Christof Schöch & Lars Schneider, Beiheft zu *Philologie im Netz* (in Vorbereitung).

Schöch, Christof & Lars Schneider (Hg., 2013). *Revolution der Medien, Evolution der Literaturwissenschaft?*, Beiheft zu *Philologie im Netz* (in Vorbereitung).

Shawe-Taylor, John, and Nello Cristianini. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge Univ. Press, 2000.

Snyman, D.P., Gerhard B van Huyssteen, and Walter Daelemans (2011). “Automatic Genre Classification for Resource Scarce Languages.” In *Proceedings of the Twenty-Second Annual Symposium of the Pattern Recognition Association of South Africa*. Vanderbijlpark, South Africa.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis (2000). “Automatic Text Categorization in Terms of Genre and Author.” *Computational Linguistics* 26/4, 471-497.

Tufte, Edward (2001). *The Visual Display of Quantitative Information*. 2nd ed. Cheshire: Graphics Press.

Witten, Ian et al. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. San Francisco: Morgan Kaufmann.

Wolf (2008), "Historische Textgattungen", in: *Rhetorik und Stilistik, Ein internationales Handbuch historischer und systematischer Forschung*, hg. von Ulla Fix, Andreas Gardt & Joachim Knape. Band 2, Berlin: de Gruyter, 1076-92.

Yang, Yiming (1999). ”An evaluation of statistical approaches to text categorization”, *Information Retrieval* 1.1, 69-90.

Zymner, Rüdiger (2003). *Gattungstheorie. Probleme und Positionen der Literaturwissenschaft*. Paderborn: Mentis.