1. Jahrestagung der Digital Humanities im deutschprachigen Raum (DHd)
Universität Passau, 25.-28.03.2014

Vortragsvorschlag

Für eine computergestützte literarische Gattungsstilistik

Christof Schöch, Steffen Pielström
(Lehrstuhl für Computerphilologie, Universität Würzburg)

Einleitung

Der vorliegende Beitrag plädiert für eine computergestützte literarische Gattungsstilistik, verstanden als eine Forschungsagenda für die Literaturwissenschaften, welche hermeneutische und quantitative Methoden verbindet. Diese Agenda wird im Zusammenhang mit einem in Vorbereitung befindlichen Forschungsprojekt zum gleichen Thema formuliert, das in der romanistischen Literaturwissenschaft angesiedelt ist. Aus diesem Forschungsprojekt werden zwei Zwischenergebnisse berichtet: das Erste betrifft die konzeptuelle Verknüpfung von Gattungstheorie und computergestützter Stilistik; das Zweite betrifft die methodische Erweiterung der Principal Component Analysis (PCA) für literaturwissenschaftliche Fragestellungen.

1. Die Agenda der computergesützten literarischen Gattungsstilistik

Die übergeordnete Zielsetzung einer computergestützten literarischen Gattungsstilistk ist es, eine tiefgehende Konvergenz herzustellen zwischen etablierten literaturwissenschaftichen Fragestellungen einerseits und quantitativen Verfahren der Textanalyse andererseits. Eine solche Konvergenz ist Voraussetzung dafür, dass sich entsprechende Forschungsvorhaben im Kernbereich der Digital Humanities ansiedeln können, in dem computergestützte Geisteswissenschaften und Angewandte Informatik nicht nebeneinander stehen, sondern sich zu einem neuen, dritten Forschungsparadigma verbinden.

Auf eine computergestützte literarische Gattungsstilistik bezogen ergeben sich daraus eine Reihe von Forschungsfragen. Einige von ihnen sind primär literaturwissenschaftlich: Wie kann die Beziehung zwischen Stil und Gattung in einer produktiven Weise konzeptualisiert werden? Wie verhalten sich Gattungsstile, Epochenstile und Autorenstile zueinander? Welche anderen Faktoren spielen für die stilistische Beschreibung literarische Texte eine Rolle? Welche automatisch identifizierbaren sprachlichen Merkmale, auf welchen Ebenen der linguistischen Beschreibung, sind Indikatoren für Gattungen? Andere Fragestellungen

stammen aus dem informatischen Bereich des *Text Mining*: Welche Verfahren der Text-Kategorisierung und des Maschinellen Lernens können eingesetzt und angepasst werden? Wie können für Verfahren wie *Support Vector Machines* die besten *kernels* definiert und geeignete *features* modelliert werden? Aus der Verbindung von literaturwissenschaftlicher und informatischer Fragestellungen ergeben sich aber auch ganz neue Fragen, die dem spezifischen Bereich der Digital Humanities zuzurechnen sind: Welche Besonderheiten natürlichsprachlicher und spezifisch literarischer Daten sind zu berücksichtigen, wenn es darum geht, möglichst generische Strategien zur Trennung von Autoren- Epochen und Gattungssignal zu entwickeln? Wie können computergestützte Verfahren so weiterentwickelt werden, dass sie einerseits auch für literatursprachliche Daten statistisch signifikant, robust und verlässlich sind, dass sie andererseits aber auch aus hermeneutischer Perspektive transparent und interpretierbar, das heißt aus literaturwissenschaftlicher Sicht bedeutungsvoll sind? Und allgemeiner, wie verändert die computergestützte Herangehensweise die Weise, wie wir über literarische Interpretation und algorithmische Analyse sowie ihre wechselseitige Beziehung nachdenken? Die Bearbeitung dieser Forschungsfragen bildet den Kern der Forschungsagenda einer computergestützten literarischen Gattungsstilistik. Zu zwei dieser Teilfragen werden hier Zwischenergebnisse berichtet.

2. Die konzeptuelle Verknüpfung von Gattungstheorie und quantitativer Stilistik

Das erste Zwischenergebnis bezieht sich auf die konzeptuelle Verknüpfung von Gattungstheorie und computergestützter Stilistik. Der Gattungsstilistik geht es um einen induktiven, deskriptiven Blick auf die stilistischen Merkmale literarischer Gattungen und Untergattungen sowie auf deren historische Entwicklung.

In der neueren Gattungstheorie hat sich die Auffassung durchgesetzt, dass literarische Gattungen sich nicht mit einem idealistischen, deduktiven Ansatz systematisieren lassen (Schaeffer 1989). 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 (Hoffmann 2009). Die Kombination mehrerer solcher "facets" definiert eine Gattung oder Untergattung, und Übergangsformen oder diachrone Entwicklungen lassen sich über den Wegfall oder das Hinzutreten einzelner "facets" erfassen.

In ähnlicher Weise wird Stil heute als ein Phänomen aufgefasst, das als „Bündel konkurrierender Merkmale“ auf unterschiedlichen linguistischen Beschreibungsebenen (Phonologie, Morphologie, Lexik/Semantik, Syntax, Plot, etc.) verstanden werden kann

(Sandig 2006, Karlgren & Cutting 1994).[1] Hier kann die computergestützte Stilistik ansetzen, denn sie ist in der Lage, induktiv und umfassend zahlreiche Merkmale - in ihrer gegenseitigen Abhängigkeit, in ihrer jeweiligen Gewichtung, und unter präziser Berücksichtigung zahlreicher möglicherweise relevanter Faktoren - zu erfassen und für die Klassifikation oder das Clustering von Texten zu nutzen. Damit wird die von Dominique Combe eingeforderte „stylistique des genres“ (Combe 2002) computergestützt realisiert. Abb. 1 fasst das Verhältnis zwischen der Theorie literarischer Gattungen und computergestützter Stilistik / Text Mining zusammen.

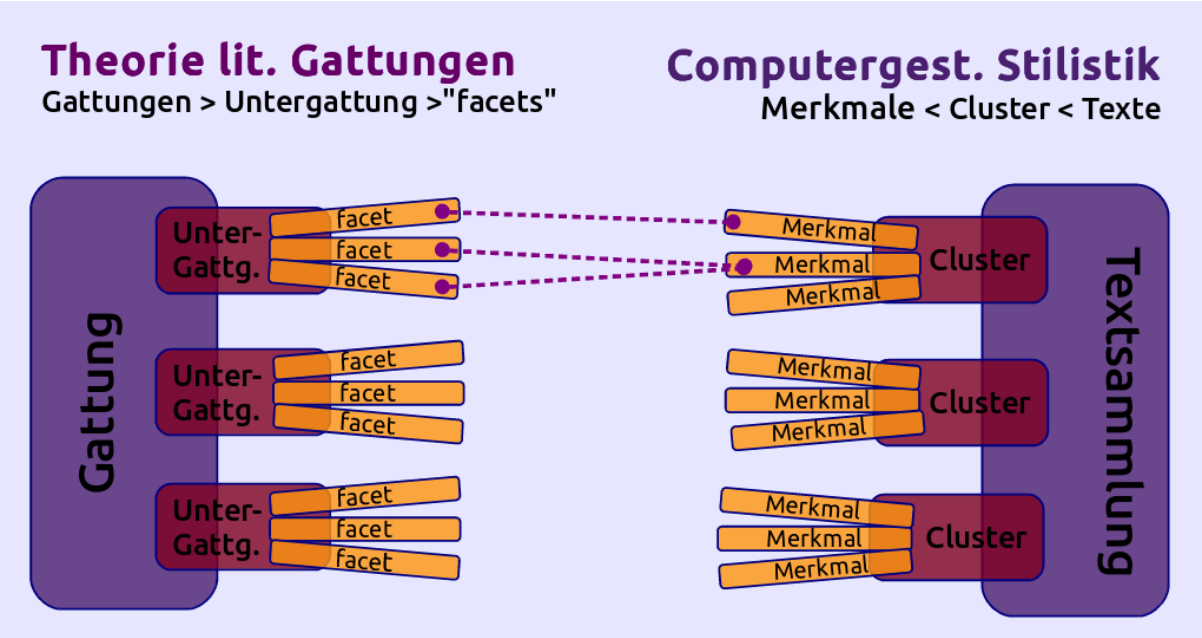

Abb. 1: Literarische Gattungstheorie und computergestützte Stilistik

Durch die vergleichbare Konzeption von Gattungen (mit Facetten) und Stil (mit Merkmalen) können Verbindungen zwischen historisch oder theoretisch gegebenen Untergattungen einerseits und auf der Grundlage stilistischer Ähnlichkeit gruppierten Clustern von Texten andererseits entdeckt werden. Genauer gesagt: es können in ihrer Stärke statistisch charakterisierbare Korrelationen zwischen einzelnen Gattungsfacetten und stilistischen Merkmalen erhoben und eingeordnet werden. Durch die Identifikation von besonders distinktiven Merkmalen und durch Merkmalsgeneralisation können dann auch

[1] 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 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 (Stamatatos et al. 2000).

Merkmalsbündel ermittelt werden, die zugleich statistisch signifikant mit einer Facette korellieren und aus literaturwissenschaftlicher Perspektive interpretierbar sind.

3. Die methodische Erweiterung der Principal Component Analysis

Die computergestützte literarische Gattungsstilistik ist Teil einer sich aktuell verstärkenden Tendenz, stilometrische Fragen über die traditionell im Vordergrund stehende Autor-Attribution hinaus zu bearbeiten.[2] Zahlreiche wohl etablierte Methoden (wie bspw. *Cluster Analysis* oder *Principal Component Analysis*), aber auch neuere informatischen Verfahren aus dem Bereich des *Text Mining* und *Machine Learning* sind für eine so konzipierte Gattungsstilistik anschlussfähig. Wir schlagen hier vor diesem Hintergrund vor, die etablierte Methode der *Principal Component Analysis* (PCA, siehe grundlegend Jackson 2005) auf eine Weise zu erweitern, die ihre Interpretierbarkeit erhöht.

Zur verlässlichen Unterscheidung von Kategorien oder Gruppen innerhalb einer großen Menge von Texten ist die Stilometrie auf die gleichzeitige Betrachtung einer großen Zahl von Merkmalen (bspw. Worten) angewiesen. Fasst man jedes Merkmal mehrerer Texte als je eine Dimensionen auf, beruht die stilometrische Erhebung von Ähnlichkeitsrelationen zwischen Texten daher häufig auf einer Dimensionsreduktion. Anders als bspw. Burrows' Delta (Burrows 2002) erlaubt PCA die Reduktion der Dimensionen eines Datensatzes nicht auf nur eine einzige, sondern auf wenige neue und voneinander unabhängige Dimensionen, die sog. *principal components* (PC), die die Varianz in den Daten besonders gut beschreiben. Jede dieser PCs korreliert hierbei mit einer spezifischen Kombination bzw. Gewichtung von Merkmalsfrequenzen. Charakteristisch für die PCA ist, daß bereits die ersten 2-4 PCs oft einen großen Teil der im Datensatz enthaltenen Varianz beschreiben (Abb. 2a). Für eine graphische Exploration der Ähnlichkeit zweier oder mehrerer Gruppen kann es also ausreichen, die ersten PCs als Koordinaten zu verwenden, anstatt in einer großen Zahl von Wortfrequenzen eine Kombination zu suchen, die Unterschiede besonders deutlich macht (Abb. 2b).

[2] Im Bereich der literarischen Gattungsstilistik liegen mehrere Ansätze vor, welche die diachrone Entwicklung von Gattungen allgemein (Moretti 2005, Jockers 2013) oder spezifische methodische Lösungsversuche betreffen: bspw. Cluster Analyse oder "unmasking"-Prozedur für die Gattungsklassifikation (Allison et al. 2011; Kestemont et al. 2012; Schöch 2013).

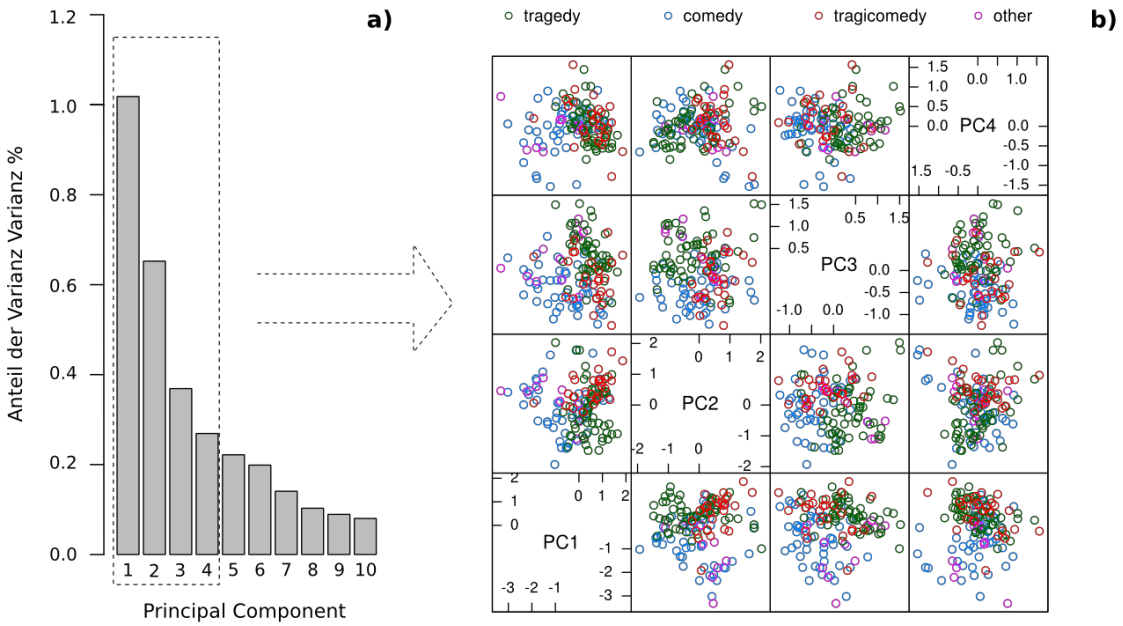

Abb. 2: Die Principal Component Analysis am Beispiel von 141 französischen Dramen (Tragödien, Komödien, Tragikomödien, Andere) aus dem siebzehnten Jahrhundert, basierend auf den relativen Häufigkeiten der 200 am meisten verwendeten Wörter. **a)** *Scree Plot* mit den Varianzanteilen für PC 1-10.; **b)** *Scatterplot Matrix* für PC 1-4 mit gattunsspezifischer Farbcodierung.

Die PCA erlaubt allerdings von sich aus keine Aussagen über die Unterschiedlichkeit zweier Kategorien von Datenpunkten, oder über den Einfluß einer bestimmten Gruppierungsvariable, weshalb stilistische Unterschiede zwischen Gattungen mit dieser Methode zwar visualisiert, aber nicht analysiert werden können. Um diese Lücke zu überbrücken, haben wir ein Verfahren entwickelt, mit dem der Einfluß der Zugehörigkeit eines bestimmten Textes zu einer spezifischen Kategorie (bspw. der Gattung) auf die PCs mittels der Varianzanalyse (ANOVA) untersucht werden kann. Hierbei wird die Gattung als unabhängige faktorielle Variable betrachtet, die Werte einer bestimmten PC als abhängige Variable. Das Bestimmtheitsmaß R^2 liefert hierbei eine Vergleichsgröße, anhand derer sich die Einflüße verschiedener Faktoren (bspw. Gattung, aber auch Autorschaft oder auch Publikationsdatum) auf eine bestimmte PC quantifizieren lassen (Abb 3a).

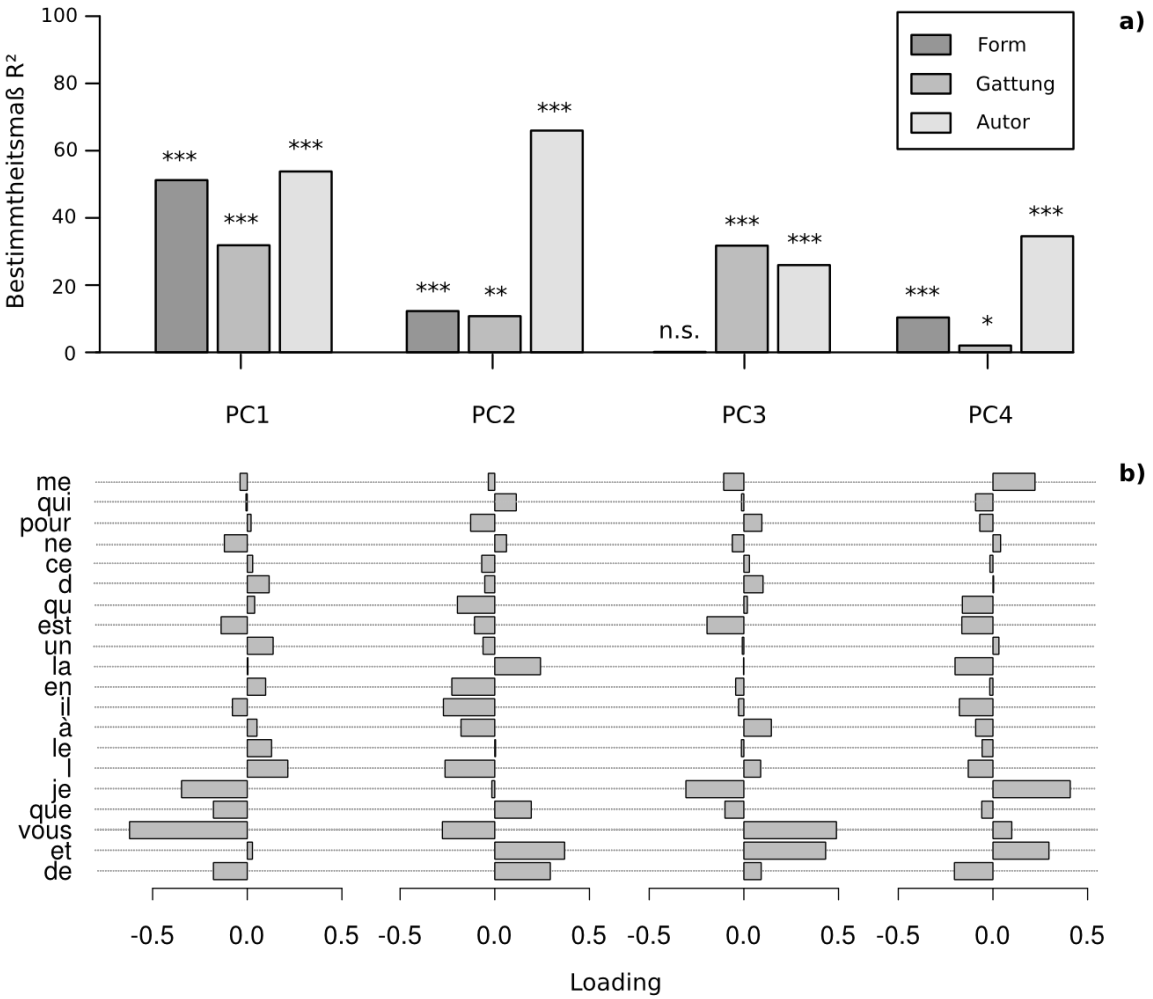

Abb. 3: Von der PCA zur Kategorie. **a)** Ergebnisse der Varianzanalysen (ANOVAs) zur Quantifizierung des Einflusses verschiedener Faktoren auf die PCs 1-4. Untersucht wurden die Faktoren Form (d.h. Vers, Prosa oder Gemischt), Gattung (Tragödie, Komödie, Tragikomödie) und Autor. Dargestellt sind die Bestimmtheitsmaße (R^2) für jeden dieser Faktoren bei jeder PC. Symbole über den Balken repräsentieren das Signifikanzniveau der jeweiligen Beziehung (‘*n.s.’* nicht signifikant; ‘*’ $p < 0.05$; ‘**’ $p < 0.01$; ‘***’ $p < 0.001$). Den stärksten Einfluß hat die literarische Gattung in diesem Datensatz auf PC1 und PC3; **b)** *Loading*-Werte der 20 häufigsten Worte für PC 1-4. Diese Werte repräsentieren den Einfluß einzelner Wortfrequenzen auf die PCs, und erlauben so in Kombination mit den Resultaten der Varianzanalysen interpretierende Rückschlüsse.

Die Kombination von PCA und ANOVA erlaubt somit die Identifikation von PCs, die durch einen bestimmten Faktor besonders stark geprägt werden, und den Abgleich mit der Gewichtung bestimmter Wortfrequenzen in dieser PC (Abb. 3b). Gemeinsam betrachtet erlaubt dies Rückschlüsse darüber, welche Merkmale und Merkmalsbündel für die Unterschiede in bestimmten Faktoren besonders relevant sind, d.h. auch welche Kombinationen charakteristisch für Gattungsunterschiede sind.

Fazit

Mit der konzeptuellen Verbindung von Gattungstheorie und computergestützter Stilistik einerseits, und der methodischen Erweiterung der PCA zur verbesserten Interpretierbarkeit von PCs in Bezug auf relevante Kategorien andererseits, konnten wichtige Zwischenziele erreicht und Grundlagen für die eingangs beschriebene Forschungsagenda gelegt werden. Und es konnte exemplarisch gezeigt werden, so hoffen wir, wie die hier vertretene Forschungsagenda einer computergestützten literarischen Gattungsstilistik hermeneutische und quantitative Methoden zu einem eigenständigen Ansatz verbindet, der auf die Entwicklung spezifisch erweiterter informatischer Verfahren für ein erweitertes Verständnis der Natur und der Entwicklung literarischer Gattungen abzielt.

Literaturangaben

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

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

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

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

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.

Jackson, Edward (2005). *A User's Guide to Principal Components*. New York: Wiley.

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.

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

Kessler, Brett, Geoffrey Numberg, and Hinrich Schütze (1998). “Automatic Detection of Text Genre.” In *Proceedings of ACL 1998*, 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.

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

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 (2013). “Fine-tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater”, *Digital Humanities Conference* 2013, http://dh2013.unl.edu/abstracts/ab-270.html.

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