ePoetics – Korpuserschließung und Visualisierung deutschsprachiger Poetiken (1770-1960) für den ‚Algorithmic Criticism‘

‚
stefan.alscher@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

‚
mbender@linglit.tu-darmstadt.de
Technische Universität Darmstadt

‚
Markus.John@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

‚
A.Mueller3@gmx.net
Universität Stuttgart, Deutschland

‚
sandra.richter@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

‚
rapp@linglit.tu-darmstadt.de
Technische Universität Darmstadt

‚
thomas.ertl@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

‚
Steffen.Koch@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

‚
jonas.kuhn@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

ePoetics ist ein Forschungskooperationsprojekt der Universität Stuttgart und der Technischen Universität Darmstadt. Gefördert vom Bundesministerium für Bildung und Forschung zielt es gleichermaßen auf einen Erkenntnisgewinn für die Informatik sowie die Sprach- und Literaturwissenschaft dank einer wechselseitigen Anregung und Ergänzung im Sinne des ‚Algorithmic Criticism‘ nach Stephen Ramsay (Ramsay 2007). Dieser Ansatz ist explizit nicht darauf ausgerichtet, lediglich hermeneutische Hypothesen mit algorithmischen Verfahren zu überprüfen. Vielmehr zielt er darauf, durch den iterativen Einsatz analoger und digitaler Methoden verschiedene Perspektiven auf Texte einnehmen und abgleichen zu können. Darüber hinaus ist ein zentraler Aspekt dieses Forschungsparadigmas, Erschließungsentscheidungen und -verfahren sowie Analyseschritte transparent bzw. nachvollziehbar und nachnutzbar zu machen. Das Projekt ePoetics ist der Digitalisierung, Annotation, Analyse und Visualisierung eines für die Geisteswissenschaften zentralen Textkorpus‘ gewidmet: Poetiken und Ästhetiken von 1770 bis 1960. Diese Texte dokumentieren das Denken und Schreiben über Literatur und andere Künste in der zentralen Periode nach der Abkehr von der Normen- und Regelpoetik (vor 1770) und vor dem Übergang zur Literaturtheorie und damit dem Ende der Poetik als literaturwissenschaftlicher Textgattung (nach 1960). Sie enthalten dabei grundlegendes Wissen über Sprache und Literatur (-wissenschaft), etwa die Erläuterungen zentraler Begriffe und deren Zusammenhänge. ePoetics betreibt die Entwicklung und Untersuchung eines Testkorpus‘ von zwanzig Poetiken, ausgewählt aus einem Gesamtkorpus von 1240 Texten (inkl. aller Auflagen), die Sandra Richter in ihrer Studie ‚A history of Poetics‘ (Richter 2010) als zur Gattung ‚Deutschsprachiger Poetik‘ zählbarer Werke bibliographiert hat. Die Auswahl des Testkorpus‘ enthält – historisch und systematisch betrachtet – die repräsentativsten Texte des Gesamtkorpus‘, d. h. die, die am häufigsten zitiert und in den meisten Auflagen herausgegeben wurden, und stellt dennoch auf den ersten Blick ein sehr heterogenes Korpus dar. Aus sprach- und literaturwissenschaftlicher Sich zeigen wir auf, wie sich diese Heterogenität im Einzelnen darstellt, aber auch, welche tiefergehenden Gemeinsamkeiten und Abhängigkeiten die Texte auf den zweiten Blick aufweisen und auf welche Ursprünge sich diese zurückführen lassen. Für ausführlichere Informationen zum ausgewählten Textkorpus und zum Projekt insgesamt besuchen Sie unsere Homepage (vgl.).

Im Zentrum unseres Interesses steht aktuell beispielsweise der Begriff der Metapher als ein zentrales sprach- und literaturwissenschaftliches Konzept, das in unserem Textkorpus verhandelt wird. Die mit diesem zusammenhängenden Fragen lauten: Wie wird der Begriff in einzelnen Poetiken verstanden und erklärt? Wie ändert sich dieses Verständnis innerhalb unseres Testkorpus‘? Welche literarischen oder theoretischen Werke werden im Zusammenhang damit genannt oder zitiert? Wie verändert sich der ‚Kanon‘ dieser Werke? Verändern sich die Zusammenhänge, in denen die Werke zitiert werden? Und

schließlich: Wie verändert sich insgesamt der Umgang mit Zitaten und deren Nachweisen?

Problemstellungen für die digitale Annotation mit dem Ziel der computergestützten Auswertbarkeit liegen bei solchen Texten und Anforderungen auf mehreren Ebenen vor: Das jeweilige Metaphernverständnis muss differenziert erschlossen und die Komponenten der Begriffsbestimmung müssen trennscharf kategorisiert werden können. Beispiele aus der Primärliteratur müssen eindeutig erkannt und den jeweiligen theoretischen Aspekten, für die sie stehen, zugeordnet werden. Und schließlich müssen die Textebenen und Referenzstrukturen der Poetik explizit gemacht werden – also wo der Autor selbst theoretisiert, wo zitiert oder paraphrasiert wird, inwiefern dies kenntlich gemacht wird oder nicht und sogar, wo bei Zitaten vom ursprünglichen Text abgewichen wird. Dies wird durch die Annotation nach einem komplexen Schema umgesetzt. Die Annotationen werden einerseits in TEI-konformen XML-Dateien publiziert, andererseits aber auch als Grundlage von computergestützten Analysen und Visualisierungen genutzt. Abbildung 1 veranschaulicht das Vorgehen im Projekt ePoetics im Sinne eines ‚Algorithmic Criticism' nach Stephen Ramsay (2007).

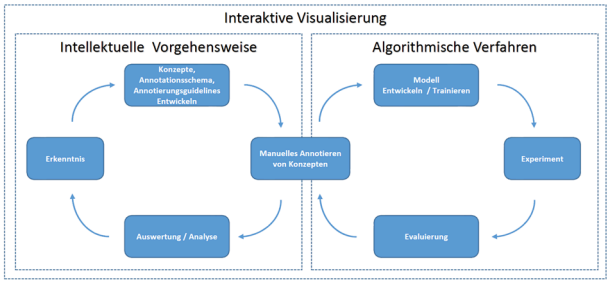

***Abb. 1**: Intellektuelle Vorgehensweise und die algorithmischen Verfahren in ePoetics, in Anlehnung an Kuhn und Reiter (2015).*

Intellektuelle Vorgehensweise (Sprach- und Literaturwissenschaft)

Die Texte des Testkorpus' stehen als Image-Digitalisate und als nach dem ‚Double Keying'-Verfahren transkribierte und aufbereitete digitale Volltexte zur Verfügung. Die strukturellen (und auch die semantischen) Annotationen des Korpus' erfolgen nach den Konventionen der Text Encoding Initiative (TEI). Das Korpus wird in virtuelle Forschungs-Infrastrukturen wie TextGrid und das Deutsche Textarchiv (DTA) integriert und dort mit den vorhandenen Referenztexten verlinkt.

Nach der Identifikation relevanter und interessanter Begriffe und Konzepte wurden zu einzelnen ausgesuchten Begriffen wie der Metapher mithilfe des UAM CorpusTool Annotationsschemata für manuelle Annotationen erstellt. Diese wurden unter ausführlicher Dokumentation von Annotationsguidelines durch mehrere Annotatoren getestet, kontinuierlich verbessert, ausgebaut und schließlich in den Poetiken durchgeführt. Abbildung 2 zeigt eine vereinfache Version des daraus hervorgegangenen Annotationsschemas, das sich in zwei Teilbereiche gliedern lässt, die teils direkt und teils mit leichten Veränderungen auch auf andere Begriffe übertragen werden können. Das Schema resultiert aus den oben genannten sprach- und literaturwissenschaftlichen Fragen, die sich in die Aspekte der Repräsentation und des Verständnisses bzw. der Anwendung des Metaphernbegriffs in den Poetiken aufteilen lassen.

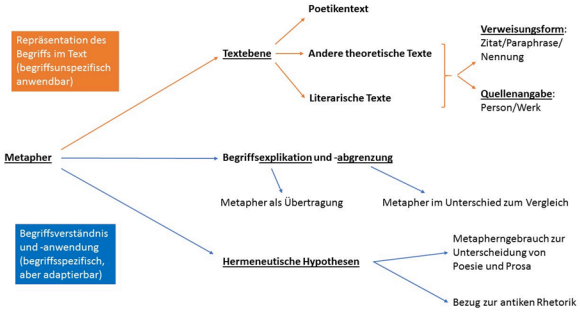

***Abb. 2**: Vereinfachte Darstellung des Annotationsschemas zur Metapher: Erkennbar ist der begriffsunspezifische (oben, orange) und -spezifische Bereich (unten, blau). Während sich der Teil des Schemas, der die Repräsentation des Begriffs im Text abbildet, sofort auf andere Begriffe anwenden lässt, ist der Teil, der sich dem Begriffsverständnis und der -anwendung widmet, begriffsspezifisch. D. h. die hier zu annotierenden Kategorien lassen sich nicht für andere Begriffe verwenden, aber leicht durch passende für den jeweiligen Begriff ersetzen.*

Das Annotationsschema stellt eine Systematisierung des Begriffs, d. h. seines Vorkommens und Verständnisses in den Poetiken dar. Die für den Begriff relevanten Textstellen werden zunächst dahingehend klassifiziert, ob es sich um Poetikentext handelt (also Text vom Autor der Poetik selbst), oder ob andere theoretische oder literarische Texte zitiert, paraphrasiert oder genannt werden. Neben den Verweisungsformen annotieren wir hierbei auch die Quellenangaben – beides im Übrigen nicht nur, wenn es explizit angegeben ist. So berücksichtigen wir auch die Möglichkeit von „versteckten" Zitaten oder solchen, bei denen die Quelle nicht oder unvollständig benannt ist. Das Auffinden bestimmter Muster sowie zum Beispiel Titel und Personennamen oder Zitate wird dabei unterstützt durch computerlinguistische Methoden und Verfahren der interaktiven Visualisierung. Darüber hinaus systematisieren wir das vorliegende Begriffsverständnis, d. h. ob die Metapher z. B. als Übertragung erklärt wird, und grenzen sie von anderen Begriffen ab, z. B. im Unterschied zum Vergleich. Zusätzlich lassen sich auch Beobachtungen zu konkreten hermeneutischen Hypothesen annotieren, z.

B. ob anhand des Metapherngebrauchs zwischen poetischer und prosaischer Sprache unterschieden wird.

Schon durch die Annotation von implizitem Wissen entsteht somit bereits bei den manuellen Annotationen eine Metaebene an Informationen, mit der der digitalisierte Poetikentext angereichert wird. Die Systematisierung erfordert eine andere Herangehensweise an den Gegenstand, als es bei einer rein hermeneutischen Analyse der Fall wäre. Ebenso führt diese zwangsläufig zur Problematisierung der Systematisierungs(un)möglichkeit eines per se komplexen, weil heterogenen Untersuchungsgegenstandes. Das Ziel der algorithmischen Weiterverarbeitung wird zum Paradigma für die systematisch-kategorisierende Ausdifferenzierung von theoretischen Begriffen, wobei diesbzgl. neue Erkenntnisse, aber auch Grenzen aufgezeigt werden können. Die Operationalisierung der Daten führt so bereits zu Erkenntnissen, bevor computertechnologische Auswertungen durchgeführt werden, womit sie sich über den Status bloßer Vorverarbeitung erheben und einen Eigenwert besitzen.

Algorithmische Verfahren (Computerlinguistik)

Mit algorithmischen Verfahren können aus kleinen Mengen annotierter Daten (aus der manuellen und damit zeitaufwendigen Annotation) große Mengen gemacht werden, indem die annotierten Arten von Informationen automatisch auf größere Datenmengen übertragen werden.

Im Folgenden wird anhand eines Beispiels in Anlehnung an den rechten Teil von Abbildung 1 beschrieben, wie die manuelle Annotation, das Training von Klassifikationsmodellen und die Analyse der Klassifikationsergebnisse ineinander greifen. Zur Klassifizierung von Text zwischen Anführungszeichen als eine der drei Klassen ‚Hervorhebung' (Wörter deren Bedeutung hervorgehoben wird), ‚Titel' (Werktitel) und ‚Zitat' (Zitate aus anderen Werken) wurde manuell ein Korpus annotiert, in dem jeder Text zwischen Anführungszeichen einer dieser drei Klassen zugewiesen wurde (Manuelles Annotieren von Konzepten). Auf der Basis dieses Korpus wurden Klassifikationsmodelle zur automatischen Erkennung dieser drei Klassen trainiert (Modell trainieren / entwickeln). Die automatischen Modelle wiederum wurden benutzt, um in anderen Poetiken Text in Anführungszeichen automatisch in diese drei Klassen einzuteilen. Es wurde durch Stichproben und formale Evaluation auf einem für diesen Zweck annotierten separaten Korpus erkannt, dass die Klassifikation gut funktioniert (Evaluierung). Da so unter anderem direkte Zitate und Werktitel automatisch erkannt werden, ermöglicht dieser Schritt wiederum die automatische Verlinkung von Werktiteln und Zitaten mit ihren Einträgen (sofern vorhanden) im TextGridRepository-Korpus (Evaluierung). Durch diese Information kann vom Analysten manuell die Verteilung von Werken und Zitaten in den Poetiken untersucht und bedeutende Werke / Zitate erkannt werden. Diese Erkenntnisse können dann wiederum als Metadaten im Dokument annotiert werden (Manuelles Annotieren von Konzepten und Metadaten).

Interaktive Visualisierung

Interaktive Visualisierung spielt eine wesentliche Rolle in der Vorgehensweise von ePoetics, siehe Abbildung 1, da sie eine zusätzliche Interaktion zwischen Forschern und den Untersuchungsgegenständen ermöglicht. Zum einen können interaktive Systeme die hermeneutischen Vorgehensweise unterstützen, indem sie den Geisteswissenschaftlern die Möglichkeiten bieten, Annotationsschemata und -guidelines zu entwerfen, Konzepte und Metadaten in Texten manuell zu annotieren sowie diese Ergebnisse zu analysieren und darzustellen. Zum anderen kann die computerlinguistische Vorgehensweise unterstützt werden, so dass Forscher Einfluss auf komplexe Prozesse nehmen können wie beispielsweise dem Trainieren maschineller Lernmethoden durch visuelle Veränderungsparameter. Durch diese Art der Interaktion kann unterstützt werden, dass Modelle mit Hilfe des Experten entwickelt, angepasst, trainiert sowie die Ergebnisse evaluiert werden können. Um diese Herausforderungen umzusetzen, wurden zwei interaktive visuelle Analysewerkzeuge konzipiert und entwickelt. Der VarifocalReader (Ertl et al. 2014), der auf einem hierarchischen Navigationskonzept basiert (Wörner / Ertl 2013), ermöglicht den Anwendern einen direkten Zugang zu Details und Dokumentquellen, während sie auf unterschiedlichen Abstraktionsebenen mit Zusammenfassungen vorhandener Annotationen interagieren können. Des Weiteren bietet das System die Möglichkeit, computerlinguistische Modelle anzupassen bzw. zu trainieren sowie Metadaten zu analysieren, zu annotieren und zu korrigieren. Eine beispielhafte Analyse ist in Abbildung 3 dargestellt, in der der Forscher einen schnellen Überblick und Zugang zur ausgewählten Annotation „Wallenstein" (in der 3. Word Cloud sichtbar) erhält.

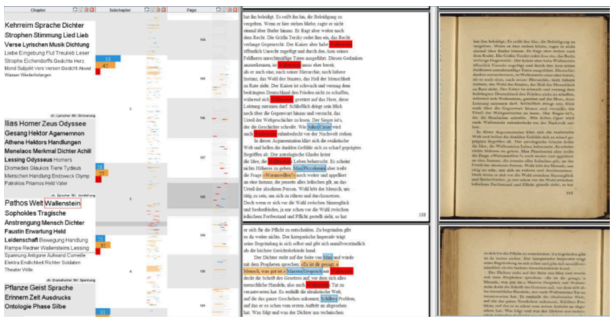

***Abb. 3**: Emil Staigers "Grundbegriffe der Poetik" unterteilt (von links nach rechts) in unterschiedliche*

Ebenen. Kapitel (mit Word Clouds), Unterkapitel (mit Balkendiagrammen und Piktogrammen), Seiten (mit Piktogrammen), Textzeilen und gescannte Digitalisate der aktuellen Seite.

Der zweite Ansatz (Heimerl et al. 2014) wurde konzipiert, um eine textvergleichende Analyse zu ermöglichen (siehe Abbildung 4). Die Visualisierung bietet einen Vergleich von mehreren Dokumenten auf einer abstrakten Ebene in Bezug auf die Verteilung der Annotationen, während die Textfelder eine flexible Navigation durch die einzelnen Texte ermöglichen. Zusätzlich unterstützt dieser focus +context Ansatz einen reibungslosen Übergang zwischen close und distant reading.

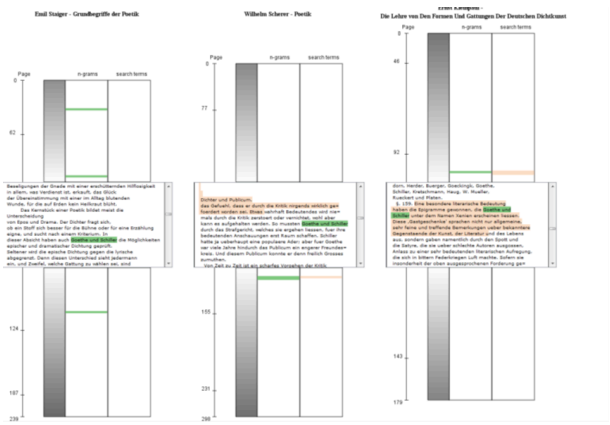

***Abb. 4**: In dieser Abbildung sind drei ausgewählte Texte nebeneinander dargestellt. Jedes dieser Dokumente verfügt über Seitenangabenskala (linke Seite) und jeweils zwei Bänder, die zum einen Annotation darstellen (grüne Balken) und zu anderen Suchergebnisse (orangene Balken). Der Analyst kann durch die einzelnen Dokumente navigieren (Textboxscrollleiste) und per Mausklick zwischen den einzelnen Annotationen (Balken) springen.*

Conclusion

Ergebnis des Projekts ePoetics ist ein digitalisiertes und annotiertes Korpus poetologischer Texte (TEI-konform und nachnutzbar), in denen zentrale Konzepte der Sprach- und Literaturtheorie durch XML-Auszeichnung explizit gemacht und systematisiert werden. Durch Korpus-übergreifende Analysen dieser Auszeichnungen können Gemeinsamkeiten und Unterschiede sowie diachrone Entwicklungen gezeigt werden. Darüber hinaus werden die Referenz- und Diskursstrukturen erschlossen (auch implizite, „versteckte“ Verweisungen), die auf verschiedenen Ebenen der Texte bestehen – einerseits Verweisungen auf andere Poetiken sowie die Identifikation bestimmter Denkschulen bzw. Theorielinien, die bis auf Ansätze aus der Antike zurückgehen (z. B. Aristoteles, Quintilian), andererseits die Diskussion von literarischen Beispielen, die Rückschlüsse auf die Entwicklungen des Literaturkanons erlauben. Die manuellen Annotationen werden iterativ gestützt durch automatisierte Methoden und Verfahren der interaktiven Visualisierung. Die dabei entwickelten computerlinguistischen Anwendungen und Visualisierungssysteme (siehe Abbildungen 3 und 4) stellen ebenfalls Ergebnisse des Projekts dar.

Bibliographie

Ertl, Thomas / Wörner, Michael (2013): “Smoothscroll. A multi-scale, multi-layer slider”, in: *Computer Vision, Imaging and Computer Graphics - Theory and Applications* 274: 142–154.

Ertl, Thomas / John, Markus / Koch, Steffen / Wörner, Michael (2014): “VarifocalReader – In-Depth Visual Analysis of Large Text Documents”, in: *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 20, 12: 1723-1732.

Ertl, Thomas / Kuhn, Jonas / Richter, Sandra / Alscher, Stefan / Rapp, Andrea (2013-2016): *ePoetics*. Universität Stuttgart [letzter Zugriff 03. Februar 2016].

Heimerl, Florian / John, Markus / Koch, Steffen / Müller, Andreas (2014): “A Visual Focus+Context Approach for Text Comparison Tasks”, in: *VisLR Workshop, LREC 2014*.

Kuhn, Jonas / Reiter, Nils (2015): “A plea for a method-driven agenda in the Digital Humanities”, in: *Proceedings of the Digital Humanities Conference, Sydney, Australia 2015*.

Ramsay, Stephen (2007): “Algorithmic Criticism”, in: Schreibman, Susan / Siemens, Ray (eds.): *A Companion to Digital Literary Studies*. Malden, MA: Blackwell 477-491.

Richter, Sandra (2010): *A History of Poetics*. German Scholarly Aesthetics and Poetics in International Context, 1770-1960. With Bibliographies by Anja Zenk, Jasmin Azazmah, Eva Jost and Sandra Richter. Berlin / New York: de Gruyter.