Tagungsbericht – Workshop: datenbankgestütztes Forschen an historischen Korpora (10.04.2018-11.04.2018)

Lisa Eggert, Lydia Dolvia, Melissa Müller

Wie kann ein historisches Korpus mit auf heutige Sprache trainierten Tools beforscht werden? Lassen sich literaturwissenschaftliche Phänomen ebenso gut in einer Datenbank abbilden wie die scheinbar eindeutigeren linguistischen Kategorien? Und wie kann eine gute Zusammenarbeit zwischen Geisteswissenschaftler*innen und Informatiker*innen funktionieren? Diesen und weiteren Fragen wollte das DFG-geförderte Projekt „Interaktionale Sprache bei Andreas Gryphius – datenbankbasiertes Arbeiten zum Dramenwerk aus linguistisch-literaturwissenschaftlicher Perspektive“ bereits in der ersten Projektphase nachgehen und hat hierzu einen Workshop veranstaltet. Die eingeladenen Teilnehmer*innen arbeiten oder arbeiteten als wissenschaftliche Mitarbeiter*innen an Projekten, die historische Korpora entweder aus linguistischer oder aus literaturwissenschaftlicher Sicht beforschen. Im Vordergrund stand dabei der Austausch der verschiedenen Projekte über konzeptionelle Entscheidungen, Arbeitsabläufe aber auch Probleme, die es noch zu lösen gilt. Deutlich wurde die enge Verzahnung fachwissenschaftlicher und informatischer Forschung, sowie die Notwendigkeit sich als datenbankbasiertes Projekt von rein editorischen Projekten abzugrenzen. Dabei wurden die Möglichkeiten aber auch Grenzen korpusanalytischer Tools herausgestellt.

LYDIA DOLIVA (Universität Duisburg-Essen) und MELISSA MÜLLER (Universität Hamburg) führten im ersten Vortrag kurz in die fachliche Ausrichtung des Gryphius-Projektes ein. Mithilfe einer in WebAnno (https://webanno.github.io/webanno/) annotierten Datenbank (auf der Basis der Datenbankarchitektur ANNIS3; http://corpus-tools.org/annis/), die das vollständige Dramenwerk von Andreas Gryphius (1616-1664) enthält, werden korpusbasiert Fragen aus literatur- und sprachwissenschaftlich übergreifender Sicht untersucht. Die Datenerhebung erfolgt auf Grundlage der historisch-kritischen Dramen-Ausgabe Eberhard Mannacks in der ‚Bibliothek deutscher Klassiker‘.
Im Gegensatz zur erprobten linguistischen Methodik sind bislang kaum literaturwissenschaftliche Verfahren zur empirischen Erhebung und Interpretation interaktionaler Sprache eingeführt. Aus diesem Grund wird in der Projektarbeit ein Set von Annotationskategorien und -kriterien neu entwickelt und kritisch reflektiert. Wesentlich sind dabei die etablierten Kategorien der Dramenanalyse, Rhetorik und Stilistik sowie der Versgestaltung. Durch die Annotation von sowohl linguistischen als auch literaturwissenschaftlichen Phänomenen können diese in der darauffolgenden Analyse miteinander kombiniert werden und neue Erkenntnisse in Bezug auf interaktionale Sprache erzeugen. So lassen sich bspw. folgende Fragen untersuchen: Korrelieren bestimmte Phänomene der Mündlichkeit (z.B. Diskursmarker, elliptische Sätze, Modalpartikeln, Anakoluthe etc.) mit den dramatischen Gattungen (Tragödie oder Komödie)? Findet die Stilisierung von Figuren auch über Formen interaktionaler Sprache statt? Und wie verhält sich die mündliche Stilisierung der Dramensprache zu der Verlaufsstruktur der szenischen Form? Die Beantwortung dieser Fragen stellen zeitgleich Ziele des Projekts dar.
Durch die Anwendung einer korpusbasierten Methodik werden aus literaturwissenschaftlicher Sicht neuartige empirische Analysemöglichkeiten erprobt, indem beispielsweise systematisch nach Regiebemerkungen, Sprecherwechseln, szenischen Konfigurationen, Versmaßen oder ausgewählten Stilfiguren gesucht werden kann.
Die sprachwissenschaftliche Ausrichtung des Projekts besteht in einer Verknüpfung von Interaktionaler Linguistik mit einer Untersuchung der Sprachstufe ‚Frühneuhochdeutsch‘ und der

Gattung ‚Drama'. Für die Beforschung interaktionaler Phänomene eignen sich die Dramen aufgrund ihrer Dialogizität in besonderem Maße. Es wird der Idee gefolgt, dass die Kenntnisse, die man empirisch anhand von Untersuchungen der heutigen gesprochenen Sprache erworben hat, in Ansätzen dabei helfen können, die gesprochene Sprache in früheren Epochen des Deutschen zu rekonstruieren. Da kein gesprochensprachliches Material aus dem Frühneuhochdeutschen vorliegt, muss auf literarische Dialoge (wie die von Gryphius) zurückgegriffen werden. Die Umsetzung der Idee findet unter Bezugnahme auf Kategorien der heutigen gesprochenen Sprache statt (Schwitalla 2006, Fiehler 2016, Hennig 2006), auf die die Tagsets bzw. die Layer der Annotation aufbauen. Darin sind z.B. enthalten: Ellipsen, (Gesprächs-)Partikeln, Responsive, Reparaturen und Vagheitsausdrücke. Das Projekt orientiert sich außerdem am korpuslinguistischen Vorgehen, das nach dem Forschungsparadigma der (historischen) Korpuspragmatik nach Felder/Müller/Vogel (2012, 5) ausgerichtet ist. Es ergeben sich unter anderem folgende Fragestellungen: Inwieweit zeigen sich in historischen, literarischen Dialogen Phänomene gesprochener Sprache, die aus dem heutigen Sprachgebrauch bekannt und beschrieben sind? Erfüllen die gesprochensprachlichen Muster wie Diskursmarker, Modalpartikeln, Ellipsen, fragmentarische Sätze, Vor-Vorfeld- und Nachfeldbesetzungen etc. die gleichen Funktionen wie in der heutigen gesprochenen Sprache oder lassen sich Differenzen feststellen? Des Weiteren soll der Frage nachgegangen werden, ob die gefundenen Phänomene Kandidaten für eine universelle Nähesprachlichkeit (gleiche Funktion wie heute) oder eher Kandidaten für eine historische Nähesprachlichkeit (andere Funktion als heute) sind (vgl. Ágel/Hennig 2006). Ebenso soll der Frage nach quantitativen Korrelationen wie oben ausgeführt nachgegangen werden.
Die technischen Aspekte, insbesondere Fragen der Datenerhebung, Annotation und Analyse, stellte MARCEL FLADRICH (Universität Hamburg) vor. Die Datenerhebung fand insofern statt, als das Dramenwerk von Andreas Gryphius abgetippt wurde, um es in einem für die Annotation und Analyse verwendbaren Format zur Verfügung zu stellen. Für die Annotation wurden die Daten in webAnno übertragen. Da ein herkömmlicher Tokenisierer für die Ansprüche des Projekts nicht ausreichend war, wurde ein individueller Tokenisierer entwickelt. Dieser nimmt die Einteilung von Sätzen zeilenbasiert vor, separiert spezifizierte Satz- und Sonderzeichen und die Daten werden als TCF-Datei gespeichert. Mithilfe des CAB (Cascaded Analysis Broker), der vom DTA (Deutsches Text Archiv) entwickelt wurde, konnten die TCF-Dateien um die Lemmatisierung, die Normalisierung und das POS-Tagging mit dem STTS erweitert werden. Um auch in webAnno ein Best-Pratice-Szenario nutzen zu können, wurden aus dem Projekt heraus drei Erweiterungen vorgenommen: 1. der Import von Orthography-Layern wurde ermöglicht, 2. Bereitstellung der Forward-Annotation für einige Layer, d.h. mit „Enter" kann innerhalb eines Layers automatisch zum nächsten Token und direkt in die Annotation gesprungen werden und 3. Definieren von festen Layerfarben zur Gruppierung und einheitlichen Kennzeichnung von Annotationslayern,. Zur Analyse der Daten werden diese dann mit Annotationen versehenen Daten aus WebAnno exportiert und in ANNIS importiert. Im Rahmen dessen wird die Entwicklung eines TSV3(.1)-Moduls für Pepper angestrebt, um auch die individuellen Annotationslayer, welche nicht im TCF-Dateiformat abgebildet werden können, aus webAnno in ANNIS übertragen zu können. Diese Modulentwicklung ist sinnvoll, um eine optimierte Konvertierung zu gewährleisten und somit schnell und einfach Annotationsergebnisse in ANNIS nutzbar zu machen.

Den Auftakt der Vorstellung anderer Projekte machte GOHAR SCHNELLE (Humboldt Universität Berlin), die in ihrem Vortrag die Annotation von Registern im althochdeutschen Referenzkorpus vorstellte. In den Mittelpunkt stellte sie die Frage, welche linguistischen Kriterien für die Erkennung von Passagen als zu einer Varietät gehörend angesetzt werden können. Eine klare Abgrenzung von Textteilen erfordert die Definition aller registerkonstituierenden Elemente, insbesondere der

Kommunikationssituation. Um sich einer Registerdefinition im Althochdeutschen zu nähern, muss jeweils eine Hypothese über die Situationseinteilung und über die sprachlichen Merkmale gebildet werden. Mit der Evangelienharmonie Otfrids von Weißenburg als Datengrundlage wurden folgende Hypothesen aufgestellt: Die Kommunikationssituation wird in den Kapiteleinteilungen ersichtlich und die sprachlichen Merkmale lassen sich durch sprachlich-prototypische Phänomene einer Narration umreißen. Der Diskurstyp „Narration" wurde hier in einer auf Clusterbildung basierenden Fallanalyse untersucht. Diese Clusteranalyse ergibt, dass 45% der hypothetischen narrativen Texte auch charakteristisch narrativ sind. Als möglich Gründe für Abweichungen wurden folgende Punkte genannt: 1. Texte sind nicht narrativ, sondern exegetisch (< 25% Präteritum), 2. Es sind Mischtexte vorhanden (> 25% Präteritum, <65% Präteritum), 3. Integrierte Kommentare sind vorhanden und 4. Texte haben eine andere, noch nicht bekannte Registerkategorie. Zusammenfassend kann festgehalten werden, dass ein Register eine situative Varietät ist und registerkonstituierende Komponenten sprachspezifische Ausprägungen haben. Dies führt dazu, dass einer Registerannotation eine sprachspezifische – hier eine althochdeutsche – Registerdefinition vorausgehen muss. Als definitorisches Merkmal sind Interferenzen zwischen linguistischer und literaturwissenschaftlicher Funktionalität in Bezug auf Erzählmodus, Fokalisierung, Rolle des Erzählers und Erzählstruktur anzunehmen.

Im zweiten Vortrag präsentierte HANNAH BUSCH (Universität Trier) das bereits abgeschlossene Projekt eCodicolgy, welches im Rahmen einer vom BMBF geförderten Kooperation (Technische Universität Darmstadt, Trier Center for Digital Humanities, Karlsruher Institut für Technologie) stattfand. In eCodicology wird anhand von Algorithmen das Tagging mittelalterlicher Handschriften erprobt. Ziel der Projektarbeit ist die automatische Erkennung von Layoutelementen auf makro- und mikrostruktureller Ebene, eine statistische Auswertung systematischer Merkmale der Kodikologie sowie die Erkennung und Visualisierung versteckter Beziehungen zwischen den Kodizes.
Bei den zu untersuchenden Parametern ist vor allem die Vermessung, Positionierung und Zahl der Elemente auf jeder Seite von Interesse. Diese werden mit Hilfe der Algorithmen automatisiert in den Handschriften untersucht und vermerkt. Durch diese Ergebnisse können später bspw. die Metadaten der Handschriften in Bibliothekskatalogen ergänzt werden.
Für die Analyse müssen die gescannten Handschriften zunächst farblich kalibriert werden, bevor in einem zweiten Schritt die Skalierung und Auflösung angepasst wird. Auf Basis dieser Daten werden dann die festgelegten Merkmale (z.B. Seitengröße, Größe des Textraums, Spaltenraums, Bildraums uvm.) identifiziert und extrahiert. Die extrahierten Daten werden als TEI-XML-Datei gespeichert und weiterverarbeitet.
Hannah Busch gab detailliert Einblicke in die Herausforderungen und Chancen einer interdisziplinären und institutsübergreifenden Projektarbeit. Grundlegend war bspw. das Finden einer gemeinsamen Kommunikationsbasis. Technische Hürden waren einerseits die Verlässlichkeit des digitalen Faksimiles in Bezug auf Farbechtheit, Auflösung, Bildrauschen, Größenverhältnis sowie Speicherformat und andererseits die heterogenen Katalogdaten der Handschriften.

Im Anschluss daran stellten SEVGI FILIZ und BERNHARD FISSENI (beide Universität Duisburg-Essen) die Arbeitsstelle Edition und Editionstechnik (AEET), ein Lehr-Forschungsprojekt, vor, das bereits seit einigen Jahren besteht. Ziel der AEET ist es Texte aus bislang unerschlossenen privaten und öffentlichen Archiven digital zu erfassen, in Datenbanken zu speichern und somit vielfältig auswertbar zu machen. Sevgi Filiz erläuterte zunächst die Arbeitsweise des Projektes, das Student*innen die Möglichkeit bietet vor Ort in den Archiven an der Erfassung und Aufbereitung der Daten mitzuwirken

und so einen guten Einblick in Editionstechniken und den Umgang mit Handschriften zu erhalten. Gleichzeitig können dank Transkriptionen, Übersetzungen und Kommentaren, die ebenfalls von Studierenden angefertigt werden, die Bestände für verschiedene Nutzer*innen zugänglich gemacht werden.
Bernhard Fisseni stellte im zweiten Teil des Vortrags die technischen Aspekte dieses Projektes vor. Am Beispiel des Archivs des Grafen von Platen, das seit 2009 von der AEET systematisch erschlossen und digitalisiert wird, zeigte er die Schritte zur Erstellung der Archiv-Datenbank auf. Eine durchsuchbare Datenbank, auf die auch von außerhalb der Universität Duisburg-Essen zugegriffen werden kann, erfüllt das Ziel der AEET, die Archive nicht nur zu erfassen, sondern auch zugänglich zu machen. Die Schwierigkeit beim Aufbau der Datenbank besteht in den teilweise recht uneinheitlichen Textdateien, in denen die Studierenden die einzelnen Archivalien in den letzten Jahren erfasst haben. Bei der technischen Umsetzung geht es einerseits um die Findung automatisierter Verfahren zur Erzeugung von Datenbank-Einträgen. Andererseits spielen konzeptionelle Überlegungen bei der Zuweisung von Metadaten eine Rolle, z.B. die Hierarchisierung von Textsorten oder die Datierung von Texten (gregorianisch oder julianisch). Die Zielvorstellung des Projektes ist eine vernetzte Datenbank, die mit Hilfe eines Thesaurus auch nach unscharfen klassifikatorischen Begriffen durchsuchbar ist.

Den folgenden Block begann FABIAN BARTELD (Ruhr-Universität Bochum) mit seiner Vorstellung von Annotationen in den Referenzkorpora Frühneuhochdeutsch und Mittelniederdeutsch/Niederrheinisch (1200-1650). Folgende Ziele der Projektarbeit wurden genannt: strukturierte Auswahl von Sprachdenkmälern, diplomatische Transkription, Annotation (PoS, Morphologie, Lemma) und die Zugänglichmachung der Texte über ANNIS. Anschließend wurde auf die Besonderheiten der Texte eingegangen, die in Abgrenzung von modernen Zeitungstexten, an denen gängige Annotationstools trainiert werden, als Nicht-Standard Texte angesehen werden. Dies zeigt sich beispielsweise an Phänomenen wie der Getrennt- und Zusammenschreibung, Interpunktion und Schreibvariationen. Die ersten beiden Phänomene wirken sich auf die manuelle Annotation und der letzte Faktor auf die automatische Annotation aus. In den Referenzkorpora wurde mit dem Annotationstool CorA gearbeitet, das erlaubt Token und Tokengrenzen zu editieren. Schreibvariationen sollten reduziert werden, da sie zum einen nicht annotierte Daten weniger gut suchbar machen und Frequenzanalysen verzerren, und sie zum anderen zu schlechteren automatischen Annotationen führen. Im ersten Schritt werden sie daher simplifiziert, sodass z.B. die verschiedenen Schreibweisen von *vorwar* (ahd. *fürwahr*: *uorwar, Uorwar, UOrwar*) zu einer Variante vereinheitlicht werden. Durch die im nächsten Schritt vorgenommene Normalisierung können Annotationstools für die Zielsprache genutzt werden. Die Normalisierung etabliert einen Standard, sodass sich z.B. *vorwar* und *uorwar* → *fürwahr* und *iu* und *iw* → *dir* ergibt. Um diese Identifikation von Schreibvarianten zu automatisieren wird auf die Kontextähnlichkeit und die String-Distanz zurückgegriffen (vgl. Barteld, Schröder und Zinsmeister 2015/2016b).
Als Fazit konnte festgehalten werden, dass nicht-normierte Schreibung sowohl die manuelle wie auch die automatische Annotation erschwert. Im Rahmen der manuellen Annotation kommt der Faktor der Getrennt-Zusammen-Schreibung zum Tragen. Dies wird mithilfe einer Kategorisierung und einer Korrektur der Segmentierung und der Transkription abgefangen. Bei der automatischen Annotation führt die Schreibvariation zu Problemen. Dem wird durch eine Simplifizierung, Normalisierung und eine Schreibvariantenidentifikation begegnet.

Unter dem Titel „Annotation als Grundlage computergestützter geisteswissenschaftlicher Forschung“ stellten SANDRA MURR und SARAH SCHULZ (beide Universität Stuttgart) zunächst CRETA (Center for

reflected text analytics) und dessen Arbeitsweisen vor. Interessant ist hier die interdisziplinäre Anlage des Großprojektes, unter dem sich kleinere Projekte aus den Bereichen Neuere deutsche Literaturwissenschaft, Mediävistik und Linguistik sowie aus der Philosophie und den Sozialwissenschaften versammeln. Auf Seiten der technischen Unterstützung gibt es Forschung zur maschinellen Sprachverarbeitung, zur Visualisierung und zu interaktiven Systemen. Im Vortrag wurde ersichtlich, inwiefern die Digital Humanities zunächst neue Methoden bereitstellen, die von den verschiedenen Fachrichtungen unterschiedlich genutzt werden können So ergeben sich aus den formal und inhaltlich heterogenen Korpora (hier im Beispiel Wertheradaptionen, Arthusromane und Bundestagsdebatten) verschiedene textanalytische Teilaufgaben, die sich aber überlappen können und teilweise von allen Fachrichtungen benötigt werden. Im Anschluss an die allgemeine Vorstellung ging Sarah Schulz auf die Entitätenannotation (Entitäten werden hier verstanden als: spezifische, durch Benennung unterscheidbare Objekte in einer echten oder fiktiven Welt) ein, die in allen drei Korpora zunächst manuell vorgenommen wird, um im zweiten Schritt die semi-automatische Annotation an den „Gold-Daten“ zu trainieren. Hierzu wird das für CRETA eigens entwickelte Annotations-Tool CRETAnno verwendet, das, anders als Tools wie webAnno, die typographischen Besonderheiten vor allem literarischer Texte abbilden kann. Sandra Murr stellte als zweites Beispiel für die Arbeit bei CRETA ihr Projekt zur Kategorisierung von sogenannten Wertheriaden, also Adaptionen von Goethes *Die Leiden des jungen Werther*, vor. Ihr Ziel ist es eine, an einem umfangreichen Korpus geschärfte Definition, des Genres Wertheriade zu erstellen, indem im ersten Schritt bereits vorgeschlagene Definitionskriterien annotiert und diese in einem zweiten Schritt in Form eines Distant-Readings analysiert werden. Zum Abschluss ihres Vortrages wiesen Sarah Schulz und Sandra Murr noch einmal auf die Wichtigkeit einer guten Kollaboration der einzelnen Disziplinen in DH-Projekten hin.

Der Vortrag zu Beginn des zweiten Workshoptages von LAURA PERLITZ (Humboldt Universität Berlin) wandte sich der Annotation von komplexen Nominalphrasen im Frühneuhochdeutschen am Beispiel des RIDGES-Korpus (Register in Diachronic German Science) zu. Dabei wurde unter anderem der Frage nachgegangen, welche Faktoren zu ambigen Strukturen innerhalb der frühneuhochdeutschen Nominalphrase führen und wie diese Ambiguitäten aufgelöst werden können. Solche Faktoren sind unter anderem: pränominale Genitivattribute, Getrenntschreibung von Komposita, definite Artikel, die an Positionen optional waren, an denen sie heute obligatorisch sind und das Auftreten von unflektierten Modifikatoren. Um eine Analyse von NN-Abfolgen (NN = normales Nomen) zu ermöglichen, wird dem Ansatz von Kopf (2016) nachgegangen, in dem aus verschiedenen Kriterien eine Unterscheidungsheuristik abgeleitet wird. Auf Grundlage dieser Heuristik werden NN-Abfolgen im RIDGES-Korpus analysiert. Es zeigt sich, dass die Kriterien nicht vollständig und auch nicht immer zutreffend sind. Infolgedessen wird ein alternatives Modell benötigt, das verschiedene Merkmale im Zusammenspiel erfasst und Ähnlichkeiten über die zugewiesenen Merkmale ermittelt. Somit kann ein Kontinuum zwischen dem, was im Neuhochdeutschen als Kompositum und dem, was als Nominalphrase mit Genitivattribut angesehen wird, dargestellt werden. Abschließend ergeben sich folgende Fragen, die weiteren Forschungsbedarf aufzeigen: Welche Merkmale werden zur Unterscheidung benötigt? Kann oder muss man die einzelnen Merkmale gewichten? Über den Einzelfall hinaus muss gefragt werden, inwiefern Wortartenkategorisierungen als Beschreibung von Ambiguitäten dienen können. Ebenso muss die Überlegung miteinbezogen werden, dass eine Betrachtung von Einzelmerkmalen optimaler sein könnte, um zu einem validen Ergebnis zu kommen.

Im folgenden Vortrag von SWANTJE WESTPFAHL (Institut für deutsche Sprache Mannheim) wurde das PoS-Tagging von Daten aus dem DFG-Projekt „Interaktionale Sprache bei Andreas Gryphius“

vorgestellt. Dabei wurde ein Tagger für Daten gesprochener Sprache genutzt und ein Ausschnitt aus dem Drama *Catharina von Georgien* als Datenbasis verwendet. Diese Passage enthält unter anderem folgende gesprochensprachliche Elemente: Interjektionen, Klitisierungen, Sprecherwechsel innerhalb von Satzstrukturen, Abbrüche und Selbstkorrektur sowie Nominalphrasen als Antwort. Des Weiteren sind Modal- und Intensitätspartikeln zu finden. Technisch wurde wie folgt vorgegangen: Um die Daten auf das Tagging vorzubereiten, wurden diese mithilfe des Transkriptionseditors FOLKER in einen Dialog transformiert und anschließend im Annotationstool OrthoNormal eine teilweise automatisierte Normalisierung vollzogen, die manuell nachkorrigiert wurde. In OrthoNormal fand dann auch das Tagging mit dem original TreeTagger Parameter-File und dem Parameter-File aus dem Projekt FOLK (Forschungs- und Lehrkorpus Gesprochenes Deutsch) statt. Auch dieses wurde manuell korrigiert. Als gemeinsame Probleme beider Tagger stellten sich heraus: die Subklassifikation von Verben, die Differenzierung von Artikeln, Relativpronomen und Demonstrativ- bzw. Personalpronomen. Des Weiteren hat sich jedoch gezeigt, dass der TreeTagger für gesprochene Sprache Modalpartikeln besser finden konnte als der originale TreeTagger. An dieser kurzen Passage aus einem Gryphius-Drama ließen sich mit Hilfe von Tools, die auf gesprochene Sprache trainiert sind, gute Ergebnisse erzielen. Dies unterstützt das Vorhaben des einladenden Projektes die interaktionale Sprache im Dramenwerk von Andreas Gryphius systematisch zu untersuchen. Dennoch ist zu beachten, dass die Annotation von Wortformen, die dem Barock bzw. dem Frühneuhochdeutschen zuzuordnen sind, u.a. aufgrund der abweichenden Orthographie, Kompositabildung und Lexik, problematisch bleiben.

Den letzten Vortrag gestalteten Marcus Willand und Nils Reiter (beide Universität Stuttgart) aus dem Forschungsprojekt QuaDramA (Quantitative Drama Analytics). Das Korpus setzt sich aus über 600 deutschsprachigen Dramen zusammen und umfasst hauptsächlich den Zeitraum von 1740 bis 1920. Mithilfe der computerlinguistischen Methode der natürlichen Sprachverarbeitung (NLP) werden die Dramen in Bezug auf Figurentypen, die Beziehungen zwischen den Figuren sowie die Entwicklung der Figuren untersucht.
So ist es möglich, anhand der Projektdaten die aktive und passive Präsenz der Hauptfiguren herauszuarbeiten, welches anhand der Dramen *Miss Sara Sampson* (1755), *Emilia Galotti* (1772) sowie der *Jungfrau von Orleans* (1801) vorgeführt wurde. Dabei kann nachgewiesen werden, dass Johanna und Miss Sara die am stärksten aktiv präsenten Figuren in ihrem jeweiligen Drama darstellen. Da dies für die namensgebende Hauptfigur zu erwarten ist, überrascht, dass Emilia in ihrem Drama nur an fünfter Stelle in Bezug auf die aktive Figurenpräsenz zu finden ist. Dieser Befund relativiert sich jedoch unter Einbezug der passiven Präsenz, welche anhand der Häufigkeit und Verteilung der Namenserwähnung der jeweiligen Figur erhoben wurde. So ist Emilia zwar im ersten und vierten Akt nicht aktiv auf der Bühne, aber dennoch passiv präsent, was mithilfe von gestapelten Säulendiagrammen und Streuungsdiagrammen anhand der Projektdaten visualisiert werden kann.
Des weiteren untersucht QuaDramA die Figurenrede wörterbuchbasiert zu den Themenfeldern Familie, Liebe, Krieg, Vernunft und Religion. Dadurch kann für die Rede bestimmter Figuren dramenintern gezeigt werden, in welchen Themenfeldern bspw. der Vater und der Liebhaber über Miss Sara reden. Anhand einer Klassifikation von Figurentypen ist es möglich, diese Befunde dramenübergreifend zu vergleichen und Aussagen über diese zu treffen: Welche Themenfelder bedient bspw. der Typus ‚Vater' und wie verändern sich diese im diachroner Perspektive innerhalb des Korpus?
Eine Herausforderung ist die Auflösung der Koreferenzen im Rahmen des NLP, da neben dem Eigennamen auch Pronomen und Nominalphrasen auf die jeweilige Figur verweisen können. Die technische Umsetzung erfolgt mit dem CorefAnnotator. Die Analyse der Koreferenzen ist wiederum

aufschlussreich für die Charakterisierung der Figuren, was Nils Reiter überzeugend an den verwendeten positiven (Sara) sowie negativen (Marwood) Nominalphrasen in *Miss Sara Sampson* darstellte.