Dramenwerkbank - Automatische Sprachverarbeitung zur Analyse von Figurenrede

,
andre.blessing@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

,
peggy.bockwinkel@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

,
nils.reiter@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

,
Marcus.Willand@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Einleitung

In diesem Beitrag stellen wir erste Einsichten aus einer quantitativen Analyse von Dramen vor, sowie unsere Konzeption für eine darauf aufbauende interaktive Werkbank, die einen Anstoß für eine Diskussion zur Tool-Unterstützung quantitativer Dramenanalyse geben soll. Die Werkbank unterstützt interessierte Forscherinnen und Forscher beim Einlesen von Dramen aus TEI-basierten Quellen und befindet sich noch in Entwicklung . Neben den in Dramen schon explizit kodierten strukturellen Informationen (Wer spricht was?) stellt die Werkbank insbesondere Möglichkeiten zur Verfügung mit Werkzeugen zur maschinellen Sprachverarbeitung auch den Inhalt der Figurenrede zu analysieren. Inspektions- und Aggregationswerkzeuge und -sichten erlauben auch die Analyse größerer Korpora.

Um die Anwendungsgebiete der Werkbank aufzuzeigen, skizzieren wir – anhand einer *Pilotstudie* zur Analyse des Verhältnisses von dramatischer Figur zur dramatischen Handlung – den Problemhorizont quantitativer Literaturwissenschaft. Dabei interessieren uns insbesondere diese Fragen: Gibt es einen Zusammenhang zwischen angenommenen prototypischen Rollen (Protagonist, Intrigant, König usw.) und Länge bzw. Häufigkeit der Redebeiträge oder der Referenz auf die Figur? Wird über bestimmte Figuren(-rollen) auf bestimmte Arten gesprochen (abwertend / aufwertend, ...)? Gibt es Figuren(-rollen)konstellationen, die häufig kookurrieren, und zwar in Bezug auf ihren eigenen Rede- und Bühnenbeitrag als auch im Bezug auf die Referenzen auf die Figuren?

Dramenanalyse: Basics

Dramentexte unterscheiden sich insbesondere durch zwei zusammenhängende Eigenschaften von Prosatexten: a) Dramatische Texte sind im Gegensatz zu vielen anderen Textsorten auf allen Ebenen (Akt- bis Redefolge) ausgesprochen gut strukturiert und ermöglichen somit eine verhältnismäßig unaufwändige Datenerhebung. Die Kehrseite der guten Strukturiertheit ist dass dramatische Texte damit nicht dem Prototyp eines Textes entsprechen, wie er von vielen Werkzeugen zur Sprachverarbeitung angenommen wird. Die maschinelle Sprachverarbeitung auf dramatischen Texten ist damit nicht durch existierende Werkzeuge „out of the box“ zu leisten. b) Die dramatischen Figuren sprechen *unvermittelt*. Unterscheidungen zwischen Erzähler- und Figurenrede und -denken spielen in Dramen keine Rolle. Während Ansätze der Stilometrie, das Figurensignal vom Erzähler- und jenes wiederum vom Gattungssignal zu trennen (Jannidis 2014), noch in den Kinderschuhen stecken, muss sich die (teil-)automatische quantitative Dramenanalyse diesen interpretativen Problemen nicht stellen. Sie hat vor allem *technisch- methodische* Probleme zu lösen: a) Erfassung und Einlesen der Daten und b) (teil-)automatische Textanalyse in Dramen. Zu letzterem gehört auch der adäquate Einsatz von interpretierbaren Maßen und transparenten Verfahren sowie visuellen Repräsentationen von Ergebnissen.

Erfassung und Einlesen der Daten: TEI-Integration

Eine automatisierte Erfassung der Oberflächenstruktur inklusiver aller relevanten Metadaten dramatischer Texte ist die Grundvoraussatzung einer quantitativen Textanalyse im oben genannten Sinne. TEI / XML ist als Standard etabliert, um Texte und Korpora möglichst genau entsprechend der/einer gedruckten Edition digital zu kodieren (cf. TextGrid; DTA). Insbesondere erlaubt TEI auch die Kodierung von Seitenzahlen, Formatierungen, Zeilenumbrüchen, Kopf- und Fußzeilen und vieles mehr, was über den reinen Textinhalt hinausgeht.

Wie Trilcke et al. (2015) auch schon festgestellt haben, ist die Extraktion der inhaltlichen Textstruktur aus den TEI-Daten keineswegs trivial. Für Netzwerkanalyse ist die eindeutige Identifizierung von Figuren besonders relevant, für eine (maschinelle, computergestützte) Analyse

des Inhaltes und der Häufigkeit der Figurenrede kommen o.g. Formatierungsmarkierungen noch als Herausforderung hinzu. In unserer Werkbank bieten wir einen Plausibilitätscheck an, der es erlaubt, Fehler im Importprozess (die sowohl durch Fehlannahmen im Importmodul als auch durch Fehlkodierungen in den Quelldaten verursacht werden können) direkt zu erkennen und zu beheben. Einmal identifizierte und behobene Fehler fließen in die Quelldaten zurück.

(Automatische) Textanalyse in Dramen

In den bereits existierenden Arbeiten zur Stilometrie auf Dramen werden komplette Dramen verglichen (z. B. durch Vorverarbeitung mit DIGIVOY). Ein differenzierter Vergleich, bei dem einzelne Figuren oder Gruppen von Figuren betrachtet werden, ist so noch nicht möglich gewesen.

Andere Projekte gehen genau den gegenteiligen Weg und verwerfen alle Dialoginhalte und beziehen ihre Netzwerkanalyse nur auf die Interaktion der jeweils in der Szene aktiven Figuren (cf. Trilcke et al.). Uns ist kein verfügbares System bekannt, das diese Lücke schließt und eine inhaltliche Analyse erlaubt, die sowohl die Interaktion der aktiven Figuren als auch deren Redeinhalt einbezieht.

In unserer Werkbank erfolgt die Textanalyse mit computerlinguistischen Werkzeugen, welche durch die CLARIN-D Infrastruktur (Mahlow et al. 2014) bereitgestellt werden. Der Aufbau von Dramen erfordert eine spezielle Herangehensweise bei der Textanalyse, da die in der Computerlinguistik oft getroffene Annahme, dass Texte aus vollständigen und grammatikalisch wohlgeformten Sätzen bestehen, in Dramen nicht zutrifft (wie auch in Texten aus sozialen Medien oder in gesprochener Sprache). Daneben weisen Dramen die oben genannte spezifische Struktur auf, die eine adäquate Vorverarbeitung bedingt. Um eine Verarbeitung mit einer nicht modifizierten CL-Verarbeitungskette zu ermöglichen, wird das Drama vorher in passende Textsegmente zerlegt. Segmente, die zu einem Dialog gehören müssen nach der Verarbeitung wieder der jeweiligen Figur zugeordnet werden. Im Kontext der Figurenanalyse sind insbesondere Eigennamenerkennung und Koreferenzresolution von Interesse. Wenn man den stilometrischen Blick weitet und auch syntaktische Konstruktionen (verwendet eine Figur mehr oder weniger komplexe Satzstruktur?) untersuchen möchte, sind auch andere linguistische Verarbeitungsschritte möglich.

Die Ergebnisse dieser Verarbeitung werden nicht fehlerfrei sein, deswegen bietet die Werkbank Möglichkeiten, die Ergebnisse zu korrigieren. Insbesondere die Zusammenführung von unterschiedlich genannten oder geschriebenen (z. B. „Emilia“ vs. „Emilie“ oder „die Soldaten“ vs. „erster Soldat“) Figuren ist nicht trivial und teilweise nur durch zusätzliches Weltwissen realisierbar. Damit dieser Schritt vereinfacht wird kommt hier ein halb-automatischer Figurenabgleich zum Einsatz. Das überarbeitete und manuell geprüfte Drama kann in einem TEI-konformen Format exportiert werden, damit die so kuratierte Ressource wieder der Community zur Verfügung gestellt werden kann. Linguistische Annotationen, die in TEI nicht direkt repräsentiert werden können, werden in einem geeigneten stand-off-Format exportiert.

Pilotstudie

In einer Pilotstudie haben wir anhand eines einzelnen Dramas exploriert, wie der Zusammenhang von (der zentralen) Dramenfigur zur dramatischen Handlung automatisiert sichtbar gemacht werden kann. Die (zentrale) Stellung im Figurennetzwerk wird dabei nicht (wie in der aktuellen Forschung gängig; vgl. Moretti 2011) lediglich durch häufige Präsenz oder Interaktion auf der Bühne repräsentiert, sondern durch differenziertere Analysen der Figurenaktivität. *Wie häufig* eine Figur spricht, *wie viel* sie spricht und *wie häufig über sie* gesprochen wird, sind dabei die Kerndaten der quantitativen Analyse, auf der weiter vorzustellende Analysen beruhen. Eine manuelle Datenerfassung übersteigt jedoch selbst bei einzelnen Dramen schnell den vom Menschen leistbaren Zeiteinsatz (wie die in Abbildung 1 manuell erstellte Erfassung der Redeteile in *Emilia Galotti* zeigt):

Figuren (Reihenfolge wie im Register)	Token von "x" im Text	Figuren-nennung	Redehäufig-keit der Figur	(Aktivitäts)-Quotient / Figur
Emilia Galotti				
emilia	126	62	64	0,45
tochter	72	72	0	
emilien	8	8	0	
emiliens	2	2	0	
emilie	1	1	0	
		145		
Odoardo Galotti				
odoardo	113	5	108	2,08
vater	47	47	0	
		52		
Claudia Galotti				
claudia	84	11	73	1,01
mutter	61	61	0	
		72		
Hettore Gonzaga, Prinz von Guastalla				
prinz	246	89	157	1,29
prinzen	33	33	0	
		122		
Marinelli, Kammerherr des Prinzen				
marinelli	301	80	221	2,76
			0	

***Abb. 1**: „ **Token von x**“ = Gesamthäufigkeit der Nennung jeder einzelnen Namensvariante. **Figurennennung** = Nennung der Namensvarianten in der Rede anderer Figuren. **Redehäufigkeit** = Wie oft spricht eine Figur. **Gesamtzahl der Wörter…** = Redelänge in Wörtern. **(Aktivitäts)Quotient** = Summe der Redehäufigkeit geteilt durch die Summe der Figurennennung: X > 1 = Aktiv (Redet häufiger als über sie geredet wird); X < 1 = Passiv.*

NLP-Unterstützte Analysemöglichkeiten in Dramen

Die Kombination von in Dramen vorhandenen strukturellen Informationen und durch automatische Verarbeitung ermittelte inhaltlich-semantische Information erlaubt neue, feinkörnige Analysen von Dramen. Die im Folgenden genannten sollen durch die Werkbank unterstützt werden, entweder durch Integration existierender oder durch Entwicklung neuer Tools.

Oberflächenanalyse der Figuren

Möglich ist eine automatische Auswertung der Figurenreden nach inhaltlichen Kriterien. Ohne Vorwissen bereitstellen zu müssen, lassen sich wichtige Begriffe, durch deren Verwendung sich eine Figur von anderen unterscheidet, mit Verfahren wie TF*IDF ermitteln und z. B. als Tabelle oder als Wortwolke darstellen. Komplexere Verfahren wie topic modeling (Blei et al. 2003) oder Wortfeldanalysen können natürlich auch auf den Redeinhalt einer Person (ggf. auf Akte / Szenen o. ä. eingeschränkt) angewendet werden, erfordern aber zumindest die Einstellung von Parametern (z. B. Anzahl der topics im topic modelling) oder das Spezifizieren von Wortfeldern. Automatische Methoden zur Erweiterung von Wortfeldern (angelehnt an z. B. Query Expansion, vgl. Manning et al. 2008) können diesen Prozess unterstützen und sollen im Rahmen der Werkbank erprobt und integriert werden. Abbildung 2 zeigt eine visuelle Auswertung dieser Analyse.

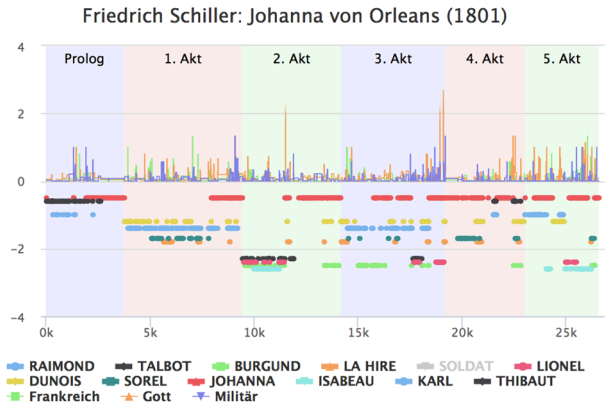

***Abb. 2**: Strukturelle und inhaltliche Analyse von Schillers Johanna von Orleans. Unten: Figurenaktivität. Oben: Prominenz ausgewählter semantischer Räume in der Figurenrede (Frankreich, Gott, Militär).*

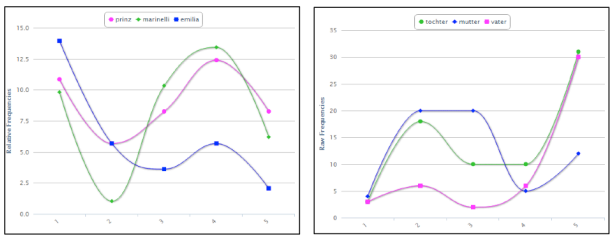

***Abb. 3**: Anhand der Häufigkeit der Figurennennung („Emilia“ vs. „Tochter“) kann der (bisher kaum erforschte) Diskursverlauf im Sinne einer Unterscheidung in private und öffentliche Konversation sehr gut nachvollzogen werden.*

Stilometrische Analysen von Figurenreden

Stilometrische Analysen werden durch eine Schnittstelle ermöglicht, durch die man Figurenrede als Datenstrukturen in R abrufen und dann nach diversen Kriterien untersuchen kann, etwa mit Hilfe von stylo (Eder et al. 2013). Es ließe sich z. B. untersuchen, ob Könige bei Schiller anders sprechen als bei Lessing, oder ob Bürgerfiguren in einem bestimmten Dramenkorpus anders sprechen als Adelsfiguren:

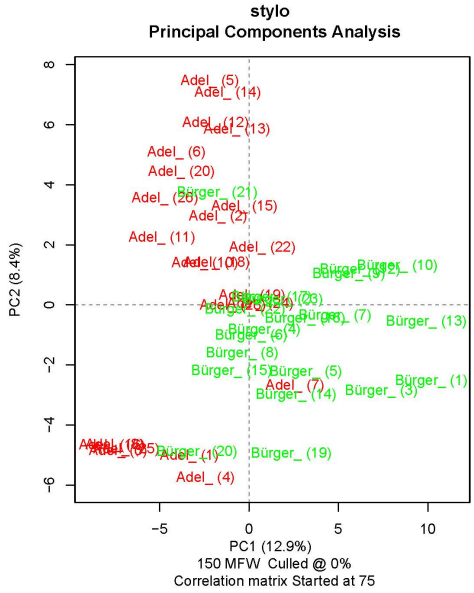

***Abb. 4**: Figurenreden, extrahiert aus 34 Dramen; nach Standeszugehörigkeit benannt.*

Sentiment-Analyse

Durch Methoden aus der Sentiment-Analyse (die zur automatisierten Analyse von Produktreviews eingesetzt wird) ließe sich z. B. analysieren, wie und ob bestimmte Figuren über andere sprechen. Neben positiv / negativ

wären auch feinere, dramenspezifische Unterscheidungen denkbar (Feigling, Hahnrei, ...).

Kombination mit Netzwerkanalyse

Die Kombination dieser Techniken mit Netzwerkanalyseverfahren würde es erlauben, im Netzwerk auch Entitäten darzustellen über die geredet wird, ohne dass sie direkt im Drama vorkommen (z. B. Gott), Kanten zwischen Knoten können dann (z. B. durch Farben) auch inhaltliche, relationale Informationen kodieren (X spricht viel / positiv über Y).

Eine Netzwerkdarstellung, in der die Position der Figuren nicht mehr zufällig (oder durch Layout-Algorithmen gesteuert) ist ist ebenfalls denkbar (Abbildung 4). Dabei werden prototypischen Figurenrollen feste Positionen in einem Raster zugewiesen, so dass große Mengen an Netzwerken schnell und direkt verglichen werden können.

Fußnoten

1. http://www.ims.uni-stuttgart.de/short/dramen

Bibliographie

Blei, David / Ng, Andrew Y. / Jordan, Michael I. (2003): „Latent Dirichlet Allocation“, in: *Journal of Machine Learning Research* 3: 993–1022.

Eder, Maciej / Kestemont, Mike / Rybicki, Jan (2013): „Stylometry with R: a suite of tools“, in: *Digital Humanities 2013 Conference Abstracts* 487-89.

Jannidis, Fotis (2014): „Der Autor ganz nah. Autorstil in Stilistik und Stilometrie“, in: Schaffrick, Matthias / Marcus Willand (eds.): *Theorien und Praktiken der Autorschaft*. Berlin: De Gruyter 169-195.

Mahlow, Cerstin / Eckart, Kerstin / Stegmann, Jens / Blessing, Andre / Thiele, Gregor / Gärtner, Markus / Kuhn, Jonas (2014): „Resources, Tools, and Applications at the CLARIN Center Stuttgart“, in: *Akten der 12. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014)* 11-21.

Moretti, Franco (2011): *Network Theory, Plot Analysis*. LiteraryLab Pamphlet 2: http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzter Zugriff 20. August 2014].

Manning, Christopher D / Raghavan, Prabhakar / Schütze, Hinrich (2008): *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Trilcke, Peer / Fischer, Frank / Kampkaspar, Dario (2015): „Digital Network Analysis of Dramatic“, in: *Digital Humanities 2015 Conference Abstracts*: http://dh2015.org/abstracts/xml/FISCHER_Frank_Digital_Network_Analysis_of_Dramati/FISCHER_Frank_Digital_Network_Analysis_of_Dramatic _Text.xml [letzter Zugriff 16. Februar 2016].