Das „Was-bisher-geschah“ von KOLIMO. Ein Update zum Korpus der literarischen Moderne

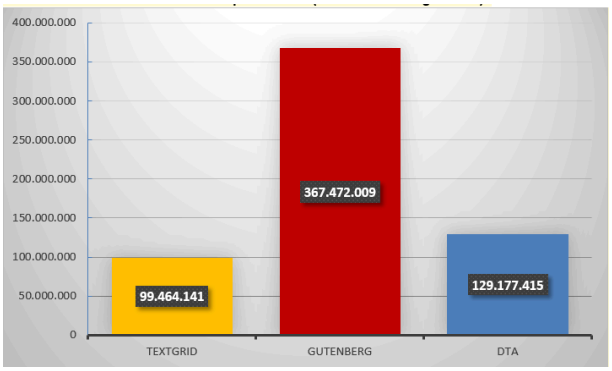
Herrmann, J. Berenike
bherrma1@gwdg.de
Universität Göttingen, Deutschland

Lauer, Gerhard
Gerhard.Lauer@phil.uni-goettingen.de
Universität Göttingen, Deutschland

Der vorgeschlagene Beitrag dokumentiert den Fortschritt beim Aufbau unseres digitalen Korpus der literarischen Moderne (KOLIMO), das im Herbst 2016 in der Beta-Version veröffentlicht werden soll (abrufbar unter https://kolimo.uni-goettingen.de/). Im Fokus des Beitrags stehen das Verfahren zur Aufbereitung der Texte (insb. Format und Metadaten in TEI) und das linguistische Tagging (POS).

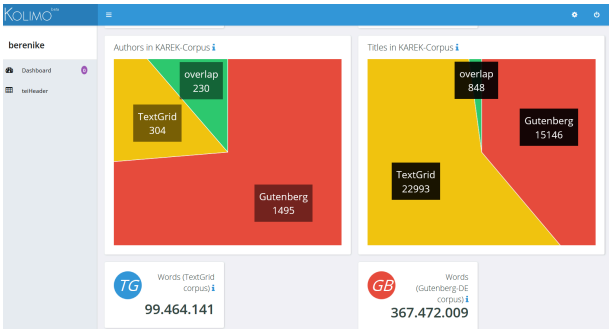
Als Teil des laufenden Projektes Q-LIMO (Quantitative Analyse der literarischen Moderne) ist KOLIMO ein repräsentatives und computerlinguistisch solide aufbereitetes Korpus von narrativen fiktionalen Erzähltexten der literarischen Epoche der Moderne. Um durch stratifiziertes Sampling Repräsentativität (verstanden als „extent to which a sample includes the full range of variability in a population“; vgl. Biber 1994) zu ermöglichen, umfasst das Korpus ein möglichst breites Spektrum der literarischen Moderne, verteilt über kanonische und nichtkanonische Texte. So wurden in das Korpus bislang ca. 596.000.000 Wörter aus frei zugänglichen Repositorien importiert (s. Abbildung 1).

Abbildung 1
Gesamtanzahl Wörter aus den drei Hauptressourcen (Zwischenstand August 2016)



Die Datenbank umfasst so neben Texten aus TextGrid und Gutenberg-DE (s. Abbildung 2) und dem DTA auch eine wachsende Zahl von Retrodigitalisaten. Das Sampling ist nicht zuletzt dadurch beeinflusst, dass KOLIMO auch das Kafka/Referenzkorpus (KAREK) beinhaltet, welches zum Ziel hat, Kafkas Texte und Texte, die Kafkas Schreibprozess beeinflusst haben könnten, möglichst umfangreich abzubilden (vgl. Herrmann / Lauer 2016a,b).

Abbildung 2
Screenshot KOLIMO-WebApp: Anzahl Wörter, Autoren und Einträge aus TextGrid & Gutenberg-DE (ohne DTA und andere Quellen, Stand August 2016)



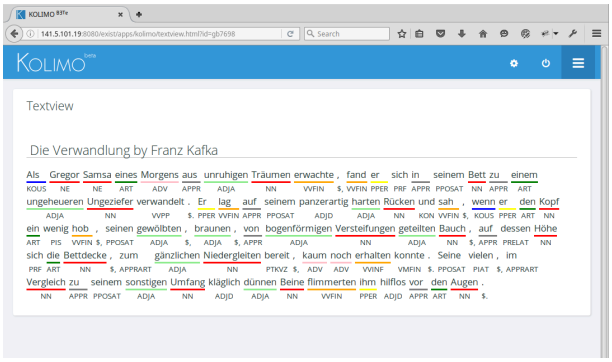
Um philologischen Ansprüchen an den editorischen Status literarischer Texte und die Abbildung von Epochen sowie Gattungskonzepten zu genügen, war eine hohe Genauigkeit und Konsistenz bei der informatischen Vorverarbeitung Textmarkup (XML-TEI) inklusive der Metadaten (Autor, Entstehungszeitpunkt und Gattung) besonders wichtig. Gerade die Auszeichnung der genannten Metadaten stellt eine Schnittstelle zwischen den informatischen und philologischen Dimensionen unseres Projektes dar: so sind Metadaten (a) die unabhängigen Variablen unserer stilistischen Analyse und (b) variieren in den von uns importierten Korpus-Ressourcen stark in qualitativer und quantitativer Hinsicht (Fehler, missing entries, unterschiedliche Ontologien). Der vorgeschlagene Beitrag wird so erstens einen kurzen Einblick in unsere Vorgehensweise geben, wobei Kriterien der Nachhaltigkeit berücksichtigt werden:

- Strategien der Textextraktion nach Genre-Kriterien unter Nutzung bestehender Metadatenschemata (ausgeschlossen wurden z.B. alle Texte, deren Metadaten sie als dramatisch und lyrisch ausflaggten, sowie Texte, die keine Absätze [without (tei:p)] enthielten);
- ein transparenter Workflow zur Korpusauszeichnung (internes eXist Webinterface);
- Anwendung eines standardisierten Text-Markups (u.a. Transformation der TextGrid und Gutenberg Header in das DTA-Basisformat TEI);
- Strategien der konsistenten Implementierung und Verbesserung von Metadatenschemata (Ineinandergreifen von händischen und

skriptgestützen Workflows, wie Recherche zu [Erst-]Erscheinungsdaten bei missing entries, Zusammenführung der unterschiedlichen Gattungschemata, Überprüfung und ggf. Zuweisung von GNDs für Autoren);
- die nachhaltige Veröffentlichung des Korpus auf einem eigenen Server mit standardisierten Datenschnittstellen;
- Datenbankabbild (nonpublic) zur Langzeitarchivierung.

Zweitens wird der Beitrag unser Vorgehen bezüglich der linguistischen Anreicherung zusammenfassen: Unter der Annahme, dass Stil quantitativ beschreibbar ist (vgl. Herrmann / van Dalen-Oskam / Schöch 2015), und dass Wortarten verlässliche Indikatoren für Register und Genrevariation sind (vgl. z.B. Biber / Conrad 2009), haben wir uns für die linguistische Annotation auf POS (STTS Tagset; vgl. Schiller / Teufel / Thielen 1995) entschieden. POS sind im Vergleich mit anderen Variationsmarkern durch eine relativ akkurate automatische Annotation besonders praktikabel. Das Webinterface liefert variablen Zugriff auf die annotierten Daten, u.a. eine Volltextansicht (siehe Abbildung 3); geplant sind zur Veröffentlichung die Exportierbarkeit in .csv-Files und TCF-Format.

Abbildung 3
Screenshot KOLIMO WebApp Textview POS-Tagging



Zwar liefern bereits trainierte Modelle von einigen Taggern (z.B. TreeTagger) eine gute Genauigkeit für das gegenwärtige Standarddeutsch, angewendet auf ältere Sprachstufen oder vom Standarddeutschen abweichende Register wie „Literatur“ sinkt die Genauigkeit jedoch. Ein bereits auf POS annotiertes Korpus ist das Deutsche Textarchiv (DTA, Berlin-Brandenburgische Akademie der Wissenschaften 2016), ein Referenzkorpus für das Deutsche, das sowohl historische Sprachstufen als auch das Register „Literatur“ enthält. Die POS-Annotation baut hier auf fehlertoleranten linguistischen Analyse historischer Texte auf und verwendet ein Tool zur Morphologisierung (Jurish 2012), ist allerdings hinsichtlich ihrer Qualität noch nicht umfassend evaluiert worden. Ausgehend von diesem Datensatz haben wir zwei Strategien verfolgt: (1) Ein epochensensitives POS-Tagging, das verschiedene Tagger auf dem Datensatz des DTA, aber auf unterschiedlichen literarischen Epochen trainiert (vgl. Paluch et al. in Vorbereitung); (2) eine Überprüfung der Qualität der DTA-POS-Tags durch quantitative und qualitative Verfahren.

In Strategie (1) machen wir uns zunutze, dass Annotationsgenauigkeit erhöht werden kann, wenn Tagger auf verschiedene Register/Sprachstände trainiert und diese trainierten Modelle dann auf noch nicht trainierte Texte des gleichen Registers angewendet werden (vgl. Giesbrecht / Evert). Für KOLIMO haben wir u.a. den TreeTagger (vgl. Schmid 1994), Perceptron (vgl. Rosenblatt 1958) und MarMoT (vgl. Müller / Schmid / Schütze 2013) verwendet. Durch die Wahl unterschiedlicher Tagger soll gewährleistet werden, dass die Genauigkeit der POS-Annotation maximiert werden kann, indem nur derjenige Tagger mit den besten Ergebnissen pro Register verwendet wird. Die Auswahl der Tagger basierte einerseits darauf, dass sie unterschiedliche Prinzipien benutzen: So funktioniert der TreeTagger nach dem Hidden Markov Model (HMM, vgl. Baum / Petrie 1966), MarMot nach dem Prinzip der Conditional Random Fields (DRF, vgl. Hammersly / Clifford 1971) und Perceptron nach dem neuronaler Netzwerke. Der Grund für die Wahl des TreeTaggers war zudem seine Prävalenz in der Forschungsliteratur, die nicht zuletzt durch gute Ergebnisse begründet scheint (vgl. Dipper 2012; Giesbrecht / Evert 2009). In einem ersten Schritt (vgl. Paluch et al. in Vorbereitung) wurden hier bereits getaggte Texte aus dem DTA in fünf Epochen geordnet. Neben der Moderne umfassten diese zu Vergleichszwecken auch Barock, Aufklärung, Romantik, und Realismus. Für die Einteilung der Epochen in Zeitperioden sowie der Einteilung von Autoren zu bestimmten Epochen wurden einschlägige Literaturgeschichten zu Rate gezogen (u.a. Beutin 2001; Jørgensen / Bohnen / Øhrgaard 1990; Meid 2009; Schulz 2000; Sprengel 1998, 2004). Anschließend wurden die Tagger auf jeweils eine Epoche trainiert, indem die Texte randomisiert in Trainings- und Evaluationstexte getrennt wurden und eine k-fold cross validation (vgl. Witten / Elbe 2005) für jeden Tagger durchgeführt wurde. Die Ergebnisse (vgl. auch Paluch et al. in Vorbereitung) weisen auf eine gute Genauigkeit insbesondere von Perceptron hin, müssen aber unter dem Vorbehalt betrachtet werden, dass der Status des DTA als Goldstandard für POS-Tagging noch fraglich ist.

Hier setzen wir mit Strategie (2) an, mit der wir zunächst für alle POS-Tags Übereinstimmung und Abweichung (Matches und Missmatches) des Outputs des Tree-Taggers und MarMots mit dem DTA-Datensatz vergleichen. Aufbauend auf diese quantitative Überprüfung der einzelnen Tag-Zuweisung evaluieren wir zudem händisch Stichproben der Nichtübereinstimmungen in der Annotation der einzelnen Tags.

Unsere quantitative Überprüfung ergibt eine generelle Übereinstimmung mit dem DTA-Datensatz in POS-Tags für den TreeTagger und den Marmot Tagger von jeweils 80%. Die generelle Übereinstimmung zwischen den Tags des TreeTaggers und denen des MarMot Taggers hingegen liegt bei 0.78%.

Tabelle 1 zeigt Ergebnisse aus der Analyse der Übereinstimmungen (Matches) und Abweichungen (Missmatches) bei der POS-Tagzuweisung von TreeTagger (TT) und MarMot (MM) im Vergleich mit den Tags des DTA. Abgebildet sind hier solche Fälle pro POS-Tag, in denen TT und MM übereinstimmen, aber vom DTA abweichen. Die Tabelle listet die elf POS-Tags, die (von TT und MM gemeinsam) die proportional den höchsten Anteil der Abweichung vom DTA ausmachen.

Tabelle 1 Abweichung zu POS-Tags des DTA (Übereinstimmung MM und TT)

POS-Tag*	Häufigkeit	Rel. Häufigkeit
NE	1444048	0.12
NN	1443795	0.12
VVFIN	1326081	0.11
ADJA	1309006	0.11
ADJD	741903	0.06
ADV	618465	0.05
VAFIN	582791	0.05
FM.la	404341	0.03
PPOSAT	397465	0.03
APPR	362774	0.03
PDAT	255896	0.02

*STTS Tagset

Aufbauend auf diesen Daten wird im nächsten Schritt die tatsächliche Qualität der bereits vorhandenen DTA-Tags für den Datensatz der literarischen Texte evaluiert. Auf der Grundlage von randomisiertem Sampling verbessern wir die POS-Annotationen bei tatsächlichen Fehlern händisch, um in der Folge u.a. eigene Sprachmodelle für unser spezifisches Korpus narrativer Texte zu trainieren. So soll schließlich unter Nutzung vorhandener Ressourcen ein Silber- oder sogar Goldstandard für das POS-Tagging historischer literarischer Texte des Deutschen erreicht werden.

KOLIMO wird in der Beta-Version zur Tagung veröffentlicht (s. https://kolimo.uni-goettingen.de) und so der Forschungsgemeinschaft zur Verfügung gestellt. Es soll eine hypothesengetriebene, aber auch explorative, quantitative Stilistik ermöglichen (vgl. Herrmann eingereicht); zum Zeitpunkt der Tagung sind erste Ergebnisse zur stilistischen Variation der literarischen Moderne zu erwarten (vgl. schon Herrmann / Lauer / Mattner 2016).

Gleichzeitig planen wir eine detaillierte Dokumentation der Arbeitsschritte zu veröffentlichen, die ähnlichen Projekten als Leitfaden zur Verfügung zu stehen soll. Unser Projekt dokumentiert in seinem gegenwärtigen Status Entscheidungen auf verschiedenen konzeptionellen, analytischen und prozeduralen Ebenen. Es zeigt, dass der Aufbau eines digitalen literarischen Korpus, das den synchronen und diachronen quantitativen Vergleich einer Schwerpunktepoche erlauben soll, bei Weitem keine triviale Aufgabe darstellt. So wurde zum Beispiel deutlich, wie Hypothesen zur Konstitution von Epochen, Autorschaft und Gattungen die Korpuskompilation steuern – und deshalb auf einer möglichst präzisen Modellierung der zugrundeliegenden textwissenschaftlichen Theorien fußen sollten. Gleichzeitig sind Metadaten (u. a. Autor, Titel, Publikationsdatum, Publikationsort, Gattung) und linguistische Parameter (wie POS) gerade die Ansatzpunkte, an denen philologische Fragestellungen in präzise und praktikable Kategorien umgewandelt werden können. Nicht zuletzt deshalb sollten literarische Daten in flexiblen Architekturen gespeichert werden, die zusätzliche Annotationsebenen zulassen – denn hermeneutische Erkenntnisprozesse stellen eine erwachsene Stärke der Geisteswissenschaften dar, die auch im digitalen Zeitalter einen explizit modellierten Platz einnehmen muss.

Bibliographie

Baum, Leonard E. / Petrie, Ted (1966): „Statistical inference for probabilistic functions of finite state markov chains“, in: *The annals of mathematical statistics 37* (6) :1554–1563.

Berlin-Brandenburgische Akademie der Wissenschaften (2016): *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/ [letzer Zugriff 24. Mai 2016].

Beutin, Wolfgang (2001): *Deutsche Literaturgeschichte: von den Anfängen bis zur Gegenwart*. Stuttgart: Metzler.

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

Dipper, Stefanie (2012): „Morphological and part-of-speech tagging of historical language data: A comparison“, in: *Workshop on Annotation of Corpora*. http://www.coli.uni-saarland.de/conf/ACRH10/slides/dipper.pdf .

Gaede, Friedrich (1971): *Humanismus, Barock, Aufklärung: Geschichte der deutschen Literatur vom 16. bis zum 18. Jahrhundert*. Bern: Francke Verlag.

Giesbrecht, Eugenie / Evert, Stefan (2009): „Is part-of-speech tagging a solved task? An evaluation of pos taggers for the German web as corpus“, in: *Proceedings of the fifth Web as Corpus Workshop* 27–35.

Hammersley, John M. / Clifford, Peter (1971): *Markov fields on finite graphs and lattices*. http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf .

Herrmann, J. Berenike (eingereicht): „In test bed with Kafka. Introducing a mixed-method approach to digital stylistics“, in: Chambers, Sally / Jones, Catherine / Kestemont, Mike / Koolen, Marijn / Zundert, Joris van (Eds.). *Special Issue DHBenelux 2015, Digital Humanities Quarterly*.

Herrmann, J. Berenike / Lauer, Gerhard (2016a): „KAREK: Building and Annotating a Kafka/Reference Corpus“, in: *DH2016: Conference Abstracts*.

Herrmann, J. Berenike / Lauer, Gerhard (2016b): „Aufbau und Annotation des Kafka/Referenzkorpus“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung*.

Herrmann, J. Berenike / Lauer, Gerhard / Mattner, Cosima (2016): *Measuring Kafka's Diaries*. A Psychostylistic Approach International Society for the Empirical Study of Literature and Media (IGEL), Chicago, USA.

Herrmann, J. Berenike / van Dalen-Oskam, Karina / Schöch, Christof (2015): „Revisiting Style, a Key Concept in Literary Studies“, in: *Journal of Literary Theory* 9 (1): 25–52.

Jørgensen, Sven Aaage / Bohnen, Klaus / Øhrgaard, Per (1990): *Aufklärung, Sturm und Drang, frühe Klassik: 1740 - 1789*. (Boor, Helmut de / Newald, Richard, eds.). München: Beck.

Jurish, Bryan (2012): *Finite-state Canonicalization Techniques for Historical German*. PhD, Universität Potsdam.

Manning, Christopher D. / Raghavan, Prabhakar / Schütze, Heinrich (2008): *Introduction to information retrieval* 1. Cambridge: Cambridge University Press.

Meid, Volker (2009): *Die deutsche Literatur im Zeitalter des Barock: vom Späthumanismus zur Frühaufklärung: 1570 - 1740*. (Boor, Helmut de / R. Newald, Richard, eds.) ([Neuausg.].). München: Beck.

Müller, Thomas / Schmid, Helmut / Schütze, Hinrich (2013): „Efficient higher-order CRFs for morphological tagging“, in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Nekula, Marek (2003): „Franz Kafkas Deutsch“, in: *Linguistik online* 13 (1) https://bop.unibe.ch/linguistik-online/article/view/879/1533 .

Paluch, Markus / Rotari, Gabriela / Steding, David / Weß, Maximilian / Moritz, Maria (in Vorbereitung): *Non-static analysis of part-of-speech tagging of historical German texts*.

Rosenblatt, Frank (1958): „The perceptron: a probabilistic model for information storage and organization in the brain“, in: *Psychological Review* 65 (6): 386.

Schiller, Anne / Teufel, Simone / Thielen, Christine (1995): „Guidelines für das Tagging deutscher Textcorpora mit STTS“, in: *Manuscript, Universities of Stuttgart and Tübingen*. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf .

Schmid, Helmut (1994): „Probabilistic part-of-speech tagging using decision trees“, in: *Proceedings of the international conference on new methods in language processing* 12: 44–49.

Schulz, Gerhard (2000): *Das Zeitalter der Französischen Revolution: 1789 - 1806*. (Boor, Helmut de / Newald, Richard, eds.) (2., neubearb. Aufl.). München: Beck.

Sprengel, Peter (1998): *Geschichte der deutschsprachigen Literatur 1870 - 1900: von der Reichsgründung bis zur Jahrhundertwende*. (Boor, Helmut de / Newald, Richard, eds.). München: Beck.

Sprengel, Peter (2004): *Geschichte der deutschsprachigen Literatur 1900 - 1918: von der Jahrhundertwende bis zum Ende des Ersten Weltkriegs*. (Boor, Helmut de / Newald, Richard, eds.). München: Beck.

Witten, Ian H. / Elbe, Frank (2005): *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann Publishers.