Aufbau und Annotation des Kafka/Referenzkorpus

,
bherrma1@gwdg.de
Universität Göttingen, Deutschland

,
gehard.lauer@phil.uni-goettingen.de
Universität Göttingen, Deutschland

Der vorgeschlagene Beitrag dokumentiert das Ineinandergreifen philologischer und informatischer Fragestellungen und Entscheidungen bei Aufbau und Aufbereitung eines digitalen Korpus für vergleichende quantitative Stilanalysen von Franz Kafkas Prosa.

In den letzten Jahren haben digitale Ressourcen wie TextGrid, das Deutsche Textarchiv [DTA], und Gutenberg-DE reichhaltige digitale Korpora von historischen Texten (literarischer und nichtliterarischer Art) zur Verfügung gestellt. Kafkas Werk selbst ist zudem fast vollständig digitalisiert. Dennoch liegen derzeit weder ein vollständiges Kafka-Kernkorpus noch ein „Kafka-Referenzkorpus“ vor, das eine sinnvolle quantitative Analyse seines Sprachgebrauchs durch den Vergleich mit ausreichend großen Stichproben anderer Texte zulässt. Unser Projekt möchte diese Lücke füllen und ein Kafka/Referenzkorpus vorstellen, das sowohl philologisch als auch informatisch solide aufbereitetet ist, und eine hypothesengetriebene aber auch explorative quantitative Stilistik ermöglicht.

Bei der Konzeption des Kafka/Referenzkorpus verfolgen wir einen autororientierten Ansatz der digitalen Stilistik. Ausgehend von der Hypothese, dass der Stil eines Autors durch von ihm rezipierte Texte beeinflussbar ist, und dass Stil quantitativ beschreibbar ist (vgl. Herrmann / van Dalen-Oskam / Schöch 2015), gehen wir zunächst vom faktischen textuellen Input Kafkas aus und ergänzen diesen durch Stichproben kanonischer und populärer zeitgenössischer Texte. Der Aufbau des Kafka/Referenzkorpus wird von drei Kriterien geleitet:

(a) Vollständigkeit von Kafkas Schriften in der Originalfassung (=Kafka-Kernkorpus);

(b) Abbildung von Texten, die Kafkas Schreibprozess beeinflusst haben könnten / Abbildung von Texten, die eine näherungsweise Repräsentativität der Epoche der klassischen Moderne herstellen (=Kafka-Referenzkorpus);

(c) eine hohe Akkuratheit bzw. Konsistenz bei informatischer Vorverarbeitung wie Normalisierung, linguistischer Annotation (*Part of Speech*), Metadaten und Textmarkup (XML-TEI) in einem *Stand-Off* Korpus, das einen hohen Grad an Forschungsflexibilität ermöglicht.

Das Kafka-Kernkorpus (je nach Zählart ca. 120 Texte) wurde dabei intern in die Dimensionen Kafka_Publikation (zu Lebzeiten/posthum) und Kafka_Genre (Literarisch, Brief, Tagebuch, Amtliche Schriften) unterteilt. Das Referenzkorpus (ca. 8.000 Texte) wurde hauptsächlich aus TextGrid, DTA, Gutenberg-DE und Gutenberg.org extrahiert, und beinhaltet Metadaten zu Autor (Name, Gender), Publikationsdatum und -Ort, sowie Gattung. Es umfasst literarische Texte der Kategorien „kanonisch“ und „populär“ ebenso wie Gebrauchstexte. Neben Kinder- und Jugendliteratur die Kafka rezipierte sind hier auch Sach- und Fachliteratur von Interesse, nicht zuletzt weil Kafkas Stil durch Elemente der Fachsprache, aber auch ein hochsprachliches „Prager Deutsch“ ohne sozio- oder dialektale Einflüsse geprägt sein soll (vgl. Nekula 2003). Zur Korpuszusammensetzung wurden Aufzeichnungen zu Kafkas Lesegewohnheiten untersucht, wobei Zeugnisse über seine Bibliothek, biographische Berichte, aber auch Dokumente zur zeitgenössischen Rezeption sowie Autor- und Werk-Indices in Literaturgeschichten konsultiert wurden (Blank 2001; Born 1990; Born / Koch 1983; Born / Mühlfeit 1979). Das Ergebnis dieses Forschungsschrittes ist eine Liste von 765 Titeln, die das Metadatum „in Kafkas Bibliothek“ tragen, und einen Schwerpunkt zu Kafkas Lebzeiten setzen, aber eben auch Werke von älteren Autoren wie Goethe und Kleist, sowie Flaubert und Dostojewski (in Übersetzung) beinhalten. Dass die von uns einbezogenen Online-Repositorien hinsichtlich der editionsphilologischen Textqualität variieren ist ein hinzunehmendes Übel, dem wir zum einen pragmatisch (Wahl der bestmöglichen verfügbaren Ausgabe; Ziel, die Fehlermarge unter 2% zu halten), zum anderen unter Hinweis auf die flexible Struktur des Korpus (Austausch durch eine qualitativ hochwertigere Version ist möglich) begegnen. Durch die nahtlose Dokumentation des Korpus wird zudem die nötige Transparenz gewährleistet um auch Nachnutzern flexible Kontrolle der Daten zu ermöglichen.

Die Hauptaufgabe der informatischen Dimension des Projektes besteht neben der Einbettung in einen praktikablen und anschlussfähigen Workflow (eXist Datenbank) und der Homogenisierung und informatischen Aufbereitung der Ausgangsdaten (Tokenisierung, Normalisierung, Lemmatisierung) in einer reliablen linguistischen Annotation auf POS (STTS Tagset). Wortarten gelten als verlässliche Indikatoren für Register und Genrevariation (vgl. z. B. Biber / Conrad 2009), und sind im Vergleich mit anderen Variationsmarkern durch eine relativ akkurate automatische Annotation besonders praktikabel. Obwohl bei der POS-Annotation gute Accuracy für das gegenwärtige Standarddeutsch mithilfe von *Hidden-Markov-Modellen* und *Markov-Modellen* erzielt wird (RF-tagger, Tree-Tagger), wurden diese Tagger auf Zeitungstexten trainiert und erfordern deshalb in unserem Korpus manuelles Fehlermanagement: Ein Ausschnitt des Gesamtkorpus (repräsentativer Querschnitt auf Satzebene, randomisiertes Sampling) wird manuell auf POS getaggt und mit dem Output der Tagger verglichen. Liegt die Übereinstimmung größer oder gleich der Standard-Accuracy (ca. 97%), ist eine umfangreiche Fehleranalyse nicht notwendig. Sollte die Accuracy

niedriger sein, wird in der Folge über ein manuell kodiertes Sample von ca. 40.000 Wörtern durch *Machine Learning* ein Algorithmus trainiert, der bessere Werte erreicht. Hierbei ist auch der Gebrauch von Taggern aus dem *Conditional Random Field* (CRF) Framework wie MarMoT vorgesehen, die eine größere Input-Spanne berücksichtigen. Der Workflow beinhaltet einen automatischen Vergleich des Tagger-Outputs durch eine eXist-Datenbank mit Annotationsinterface. Der Output wird in einem Stand-Off Format (TCF) gespeichert, wie es auch das DTA benutzt. Geplant ist zusätzlich eine Qualitätskontrolle des TEI-Markups, der Metadaten und der POS-Annotation durch einen bereits entwickelten Ansatz der *Gamification* (s. https://personae.gcdh.de/index.html). Das Kafka / Referenzkorpus soll im Rahmen der TextGrid Infrastruktur in SADE veröffentlicht und so der Forschungsgemeinschaft zur Verfügung gestellt werden. 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 Autor-Korpus, das den quantitativen Vergleich mit synchronen und diachronen Daten erlauben soll, bei Weitem keine triviale Aufgabe darstellt. So wird zum Beispiel deutlich, wie Forschungsfragen beziehungsweise Hypothesen zur Konstitution von Schreibweisen und Autorschaft 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, Genre) 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

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

Blank, Herbert (2001): *In Kafkas Bibliothek*. Werke der Weltliteratur und Geschichte in der Edition, wie sie Kafka besaß oder kannte; kommentiert mit Zitaten aus seinen Briefen und Tagebüchern. Stuttgart: Blank.

Born, Jürgen (1990): *Kafkas Bibliothek*. Ein beschreibendes Verzeichnis; mit einem Index aller in Kafkas Schriften erwähnten Bücher, Zeitschriften und Zeitschriftenbeiträge. Frankfurt am Main: S. Fischer.

Born, Jürgen / Koch, Elke (eds.) (1983): *Franz Kafka: Kritik und Rezeption, 1924-1938*. Frankfurt am Main: S. Fischer.

Born, Jürgen / Mühlfeit, Herbert (eds.) (1979): *Franz Kafka: Kritik und Rezeption zu seinen Lebzeiten, 1912-1924*. Frankfurt am Main: S. Fischer.

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.

Nekula, Marek (2003): "Franz Kafkas Deutsch", in: *Linguistik online* 13, 1 https://bop.unibe.ch/linguistik-online/article/view/879/1533 [letzter Zugriff 29. Dezember 2015].