Olav Mueller-Reichau

Čechov digital – Bericht von einer Gratwanderung

Aufbruch

Philologische Institute sind traditionell auf zwei tragenden Säulen gebaut, der Literaturwissenschaft und der Linguistik. So auch unser Institut für Slavistik in Leipzig. Beide Fachbereiche haben ihre jeweils eigene Entwicklung genommen, und man kann wohl mit Recht sagen, dass sie sich heute, wenn überhaupt, dann nur selten gegenseitig bereichern. Es ist offensichtlich sehr schwer, interessante Forschungsfragen zu stellen, deren Beantwortung von der Expertise der Literaturwissenschaft und der Expertise der Linguistik gleichermaßen profitiert.

Der vorliegende Artikel beschreibt den Aufbruch in eine Unternehmung, die sich die Entwicklung und Verfolgung einer solchen Fragestellung zum Ziel gesetzt hat. Ich werde ein laufendes Projekt vorstellen, das ich zusammen mit Matthias Irmer (OntoChem GmbH, Halle/Saale), Anastasija Koretskykh (Institut für Slavistik, Universität Leipzig), Michael Richter (Institut für Informatik, Universität Leipzig) und Tariq Yousef (Institut für Informatik, Universität Leipzig) durchführe. Das Phänomen, dem wir uns verschrieben haben, ist ein klassisch literaturwissenschaftliches, das in den Kompetenzbereich der Jubilarin fällt: der Subtext im erzählerischen Werk von A. P. Čechov. Der Blick, den wir auf den Subtext haben, ist ein linguistischer: wir verstehen darunter den pragmatischen Bedeutungsanteil einer Äußerung, der den semantischen Bedeutungsanteil vervollständigt (dazu später mehr). Die Methode, mit der wir dem Subtext zu Leibe rücken wollen, ist eine digitale Textanalyse, fällt also in den Kompetenzbereich der Informatik.

Über den Subtext im Werk Čechovs ist viel Wissen zusammengetragen worden.[1] Es versteht sich von selbst, dass ich nicht behaupten werde, die Fülle der Literatur auch nur im Ansatz zu überblicken. Stattdessen wage ich nur ein paar allgemeine Feststellungen.

Čechovsche Erzählungen sind in der russischen Standardsprache verfasst. Sprachliche Extravaganz steht ihnen fern. Für die direkte Rede einzelner

1 Für einen Überblick vgl. Elena I. Lelis: Podtekst kak lingvopoėstetičeskaja kategorija v proze A. P. Čechova. Iževsk 2013.

Charaktere im Werk mag teilweise anderes gelten, die Erzählersprache zumindest ist in Lexik und Syntax grundsätzlich konservativ.[2] Aufmerksamkeit erzielen Čechovs Texte also nicht durch besonderes sprachliches Material, lexikalische Innovationen etwa oder andere ungewöhnliche Sprachformen.

Die ›Action‹ findet bei Čechov nicht an der Textoberfläche statt – dieses Gefühl hat vielfältig Ausdruck erfahren. Mit Bezug auf das dramatische Werk zum Beispiel in der folgenden Theaterkritik:

> Tim Kramer verzichtet am Theater St. Gallen auf jede politische oder sonstwie zeitgenössische Anspielung. Vermutlich ist das klug, weil Tschechows Diagnose zwar aus der Zeit geschrieben, jedoch nicht an Zeit und Ort gebunden ist. Sein »Moskau« ist Chiffre für das ewig Unerfüllbare, das die Menschen seit je und bis heute umtreibt. Dies allerdings würde man gern »aktualisiert« sehen, nämlich in packenden Figuren, in zugespitzten Konflikten, wie sie der Menschenkenner Tschechow mit Sätzen erschafft, in denen äusserlich nichts passiert und innerlich Welten zusammenbrechen.[3]

Sätze zu erschaffen, in denen äußerlich nichts passiert und innerlich Welten zusammenbrechen – diese künstlerische Leistung fasziniert. Ich versuche einmal, sie linguistisch einzuordnen. Sprache ist, wie spätestens seit de Saussure allgemein bekannt, ein Zeichensystem, das seiner Natur nach Formen und Bedeutungen paart. Der Bereich des ›Äußerlichen‹ ist durch die äußere sprachliche Form abgesteckt. Was durch explizite Formen beschrieben wird, das »passiert in Sätzen äußerlich«. Daneben passiert in Sätzen aber auch etwas »innerlich«, und bei Čechov offenbar sehr viel.

Aber wie ist es möglich, dass mittels Sprache, als Zeichensystem verstanden, etwas ohne äußere Form beschrieben wird? Es muss doch etwas geben, das die Bedeutung trägt! Kein ›signifié‹ ohne ›signifiant‹! Weil es, im Rahmen des saussureschen Zeichens gedacht, einen materiellen Bedeutungsträger *per definitionem* geben muss, aber keiner zu sehen ist, kommt unweigerlich der Gedanke an eine unsichtbare Form auf.

Unsichtbare Materie: Hinrichs beobachtet, wie der ursprünglich aus der Astronomie stammende Begriff der ›Dunklen Materie‹ (in teilweise variierenden Formulierungen) in andere Wissenschaftsbereiche Einzug hält.[4] Der

---

2 Rolf-Dieter Kluge: Anton P. Čechov. Eine Einführung in Leben und Werk. Darmstadt 1995. S. 49.

3 Der Rezensent Peter Surber über eine Inszenierung der *Drei Schwestern* in St. Gallen. URL: saiten.ch/ticket-nach-moskau (9. April 2014).

4 Uwe Hinrichs: Die Dunkle Materie des Wissens. Über Leerstellen wissenschaftlicher Erkenntnis. Gießen 2014.

Subtext ist ein Paradebeispiel für literarische dunkle Materie. Den Subtext durch geschickte Gestaltung des Oberflächentexts zielgenau ins Bewusstsein der Leserin bzw. des Lesers zu lenken, darin besteht die Meisterschaft: »Jede Art der Interpretation des Gelesenen und der Akte des Lesens ist eine Verwandlung von dunkler in helle Materie«.[5]

Tatsächlich ist die Entdeckung dunkler Materie in Texten aus linguistischer Sicht keine Überraschung. Zumindest nicht für die Linguistik, die von einer Trennung zwischen semantischen (sprachlichen) und pragmatischen (außersprachlichen) Bedeutungen ausgeht, welche gemeinsam die Bedeutung einer sprachlichen Äußerung füttern. ›Semantische‹ Bedeutungsanteile an der kommunizierten Botschaft sind demnach all die Informationen, die durch die äußere sprachliche Form kodiert sind. ›Pragmatische‹ Bedeutungsanteile sind all jene, die durch den Gebrauch der Form im Äußerungskontext impliziert werden. Das »Äußerliche« im obigen Zitat entspricht also der semantisch kodierten Bedeutung, das »Innerliche« der pragmatisch inferierten.

Bevor ich beschreibe, wie wir versuchen wollen, der dunklen literarischen Materie bei Čechov habhaft zu werden, erlaube ich mir ein paar einordnende Bemerkungen zu der Methode, die wir verfolgen.

Kann man Bedeutungen messen?

Bedeutungen kann man natürlich nicht messen. Sie haben kein Gewicht und keine Extension im Raum. Sie können in Vergessenheit geraten, aber sie können nicht verschwinden. Bedeutungen sind am ehesten so etwas wie erinnerte Gefühle, die (re)aktiviert werden, wenn man sie triggert. Ein wichtiger Bedeutungstrigger sind Wörter. Weil sich die Linguistik mit der menschlichen Sprache auseinandersetzt, gehört die Erforschung sprachlicher Bedeutungen zu ihren Kernaufgaben. Hier stellt sich die Grundsatzfrage, wie man Bedeutungen überhaupt erforschen kann, wenn sie nicht messbar sind.

Jener Teilbereich der Linguistik, der sich die Erforschung von sprachlichen Bedeutungen ausdrücklich zum Ziel macht, ist die Semantik. In der Semantik werden, soweit ich sehe, grundsätzlich zwei Wege verfolgt, um der Bedeutungen habhaft zu werden. Beiden Ansätzen ist gemeinsam, dass sie Bedeutungen nicht direkt unter die Lupe nehmen. Das geht ja auch gar nicht, denn Bedeutungen sind nicht messbar. Stattdessen wendet man sich Effekten zu, die Bedeutungen haben.

5 Uwe Hinrichs: Dunkle Materie. S. 152.

Der erste Weg besteht darin, Bedeutungen in Aktivierungszustände des Gehirns zu übersetzen. Verschiedene Bedeutungen entsprechen dann verschiedenen neuronalen Aktivierungsmustern. Fällt ein Wort, so löst dies einen bestimmten mentalen Zustand aus, und diesen Zustand kann man, im Prinzip wenigstens, messen. Bedeutungen werden zu mentalen Individuen (›units‹), zu kognitiven Routinen.

Wer diesen Weg verfolgt, korreliert Wort- und Satzbedeutungen mit mentalen Repräsentationen. Entsprechende Semantikschulen entwickeln Modelle der mentalen Repräsentation. Ein prominentes Beispiel unter vielen anderen ist die kognitive Grammatik von Langacker.[6]

Der zweite Weg besteht darin, Bedeutungen in Wahrheitsbedingungen zu übersetzen. Ausgangspunkt ist die Überlegung, dass (Deklarativ-)Sätze geäußert werden, um Wahrheiten zu behaupten. Aber darf man sich heute, in postmodernen Zeiten, noch auf Wahrheiten beziehen? Das muss man sogar. Denn sprachliches Handeln, zwischenmenschliche Kommunikation, basiert auf der Prämisse, dass die Sprecherinnen und Sprecher das glauben, was sie anderen gegenüber behaupten. Grice hat diese Einsicht im Rahmen seines bekannten Kooperationsprinzips als Kommunikationsmaxime der Qualität formuliert: »Try to make your contribution one that is true«.[7]

Wenn mir jemand etwas mitteilt, muss ich grundsätzlich davon ausgehen, dass diese Person eine Wahrheit behauptet. Ich muss diese behauptete Wahrheit natürlich nicht als Wahrheit akzeptieren (»Nein, du irrst dich!«), aber ich muss stets davon ausgehen, dass die Botschaft als Wahrheit gemeint ist. Sonst funktioniert Kommunikation nicht.[8]

Wenn geäußerte Sätze behauptete Wahrheiten sind, ergibt sich daraus die Möglichkeit, indirekt an die Bedeutung des Satzes heranzukommen. Davidson hat diesen Gedanken, auf Ideen von Frege und Tarski aufbauend, in die Semantik eingeführt.[9] Das Credo der wahrheitskonditionalen Semantik lautet

---

6 Ronald W. Langacker: Foundations of Cognitive Grammar. Bd. 1. [Theoretical Prerequisites]. Stanford 1987; Ronald W. Langacker: Foundations of Cognitive Grammar. Bd. 2. [Descriptive Application]. Stanford 1991. Ronald W. Langacker: Cognitive Grammar: A Basic Introduction. Oxford 2008.

7 Herbert P. Grice: Studies in the Way of Words. Cambridge 1989. S. 27. Zu der Maxime der Qualität gehören bei Grice zwei untergeordnete Maximen: »Do not say what you believe is false« und »Do not say that for which you lack adequate evidence«.

8 Eine Lüge wäre keine Lüge, wenn sie nicht vor dem Hintergrund der Erwartung einer Wahrheitsbehauptung beurteilt werden würde. Ebenso ist das Phänomen der Ironie nur als Spiel mit der erwarteten Wahrheit verstehbar.

9 Donald Davidson: Truth and Meaning. In: Synthese 17/3 (1967). S. 304–323.

entsprechend: »Die Bedeutung eines Satzes zu kennen, heißt, zu wissen, wie die Welt beschaffen sein muß, damit der Satz wahr (oder falsch) ist«.[10]

Während die kognitive Semantik Bedeutungen mit internen (kognitiven) Zuständen korreliert, korreliert die wahrheitskonditionale Semantik Bedeutungen also mit externen (Welt-) Zuständen. Nun hat sich im Zuge der rasanten Entwicklung digitaler Ressourcen im Bereich der sog. ›digital humanities‹ ein dritter Weg aufgetan, wie man Bedeutungseffekte messen kann. Diesen Weg beschreiten wir im Projekt.

Shannon-Information

Computer ermöglichen es, mit einer riesigen Menge von Sprachdaten (›big data‹) umzugehen. Daraus ergibt sich eine neue Idee, wie sich Bedeutungen operationalisierbar machen lassen. Die Grundannahme ist dabei, dass die Bedeutung eines Wortes die Wahrscheinlichkeit beeinflusst, mit welchen anderen Wörtern das Wort zusammen in einem Text auftritt. Wenn man für ein Wort also über eine hinreichend große Menge von Textstellen, in denen es erscheint, erfasst, mit welchen anderen Wörtern es erscheint, dann erhält man daraus ein für das Wort charakteristisches Profil. Dieses Profil, der ›Kotext‹ des Wortes, sollte dann eine Funktion seiner Bedeutung sein. Oder anders gesagt: Die Bedeutung eines Wortes lässt sich zwar nicht direkt messen (s. o.), aber es erscheint möglich, den Effekt zu messen, den die Bedeutung des Wortes auf seinen Kotext hat.

Diese Idee liegt auch dem von einer Arbeitsgruppe um Michel Richter und Tariq Yousef an der Universität Leipzig entwickelten Topik-Kontext-Modell zugrunde.[11] Das Topik-Kontext-Modell (TCM) berechnet den Informationsgehalt eines Wortes, gegeben eine Menge von Topiks innerhalb des Textes, in dem das Wort erscheint. Der Informationsgehalt, auch Shannon-Information

10 Horst Lohnstein: Formale Semantik und Natürliche Sprache. Berlin u. a. 2011. S. 2.

11 Max Kölbl, Yuki Kuogoku, Nathanael Philipp, Michael Richter, Christoph Rietdorf, Tariq Yousef: Keyword Extraction in German: Information-theory vs. Deep Learning. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020). Bd. 1/2 (2020). S. 459–464. Max Kölbl, Yuki Kuogoku, Nathanael Philipp, Michael Richter, Christoph Rietdorf, Tariq Yousef: The Semantic Level of Shannon Information: Are High Informative Words Good Keywords? A Study on German. In: R. Loukanova (Hg.): Natural Language Processing in Artificial Intelligence. Cham 2021. S. 467–491. Michael Richter, Tariq Yousef: Information from topic contexts: the prediction of aspectual coding of verbs in Russian. In: SIGTYP 2020: The Second Workshop on Computational Research in Linguistic Typology (ACL Special Interest Group on Typology). Workshop at EMNLP 2020.

genannt, ist – in Anwendung auf linguistische Fragestellungen und etwas lässig gesprochen – der Grad der Überraschung, mit dem ein Wort im Text erscheint.[12] Der Informationsgehalt ist also nicht die Bedeutung des Wortes, sondern abermals ein Effekt der Bedeutung des Wortes. Diesmal nicht sein Effekt im Sinne des kognitiven Zustands, den das Wort auslöst. Und auch nicht sein Effekt im Sinne seines Beitrags zum Zustand der Welt, der mit dem Satz, in dem das Wort erscheint, behauptet wird. Sondern sein Effekt in Sinne des Zustands des Texts, in dem das Wort erscheint.

Voraussetzung für die Berechnung des Informationsgehalts eines Wortes in einem Text ist im Rahmen des TCM die Kenntnis über die Topiks, die der Text enthält. Unter ›Topik‹ wird ganz allgemein ein Thema verstanden, um das es in einem Textdokument geht. Ein Text kann natürlich, je nachdem, wie grob- oder feinkörnig man ihn betrachtet, von mehr oder weniger vielen Themen handeln. Welche das sind, das wird durch das Verfahren der Topikmodellierung (›topic modelling‹) ermittelt. Durch Topikmodellierung wird bestimmt: Welche Topiks kommen im gesamten Textdokument (globale Umgebung) vor? Wie stark sind diese Topiks in dem Textteil (lokale Umgebung), in dem das Zielwort erscheint, präsent?

Wir benutzen zur Topikmodellierung den Algorithmus LDA.[13] LDA steht für ›Latent Dirichlet Allocation‹. Dabei handelt es sich um eine unüberwachte Technik des maschinellen Lernens, die »auf einer wiederholt zufälligen Auswahl an Textsegmenten [basiert], wobei innerhalb dieser Segmente jeweils die statistische Häufung von Wortgruppen erfasst wird«[14]. LDA basiert auf der Annahme, dass (i) jedes Textdokument eine statistische Mischung von Topiks enthält, dass (ii) ähnliche Topiks mit ähnlichen Wahrscheinlichkeitsverteilungen von Wörtern korrelieren und dass folglich (iii) jedes Topik durch eine spezifische Wahrscheinlichkeitsverteilung der Wörter charakterisierbar ist.

12 Roger Levy, T. Florian Jaeger: Speakers optimize information density through syntactic reduction. In: Bernhard Schlökopf, John Plat & Thomas Hoffmann (Hg.): Advances in Neural Information Processing Systems (NIPS) 19. Cambridge, MA, 2007. S. 849–856.

13 David M. Blei, Andrew Y. Ng and Michael I. Jordan: Latent Dirichlet Allocation. In: Journal of Machine Learning Research 3 (2003). S. 993–1022.

14 Jan Horstmann: Topic Modeling. In: forTEXT. Literatur digital erforschen (2018). URL: https://fortext.net/routinen/methoden/topic-modeling (9. Juni 2021).

Der Originaltext

Als globale Umgebung wählen wir eine Erzählung von Čechov. Da das statistische Verfahren einen gewissen Textumfang zur Voraussetzung hat, wählen wir eine relativ lange Erzählung: *Palata No. 6*. Als lokale Umgebungen wählen wir einzelne Paragraphen. Ein Dialog zwischen zwei Paragraphen wird ebenfalls als eine lokale Umgebung aufgenommen. Auf diese Weise zergliedert sich das Textkorpus (= der Gesamttext der Erzählung *Palata No. 6*, globale Umgebung) in 180 Textdokumente (= lokale Umgebungen).

Mittels LDA lassen wir berechnen: Erstens 20 Topiks, die der Gesamttext enthält. Zweitens einen Wert für jeden der 20 Topiks in jedem der 180 lokalen Umgebungen (nicht selten ist der Wert für ein Topik in einer lokalen Umgebung gleich Null, das heißt, das Topik kommt in diesem Textabschnitt gar nicht vor). Sobald wir diese Zahlen aus der Topikmodellierung erhalten haben, können wir den nächsten Schritt machen und im Rückgriff auf das Topik-Kontext-Modell für jedes Wort in der Erzählung dessen Informationsgehalt berechnen, also seinen ›Überraschungsgrad‹. Es gilt: je überraschender ein Wort, desto informativer ist es.[15]

Jedes der 20 Topiks besitzt für jede der lokalen Umgebungen einen Wert zwischen 0 und 1. Jetzt fließen alle Topiks mit einem Wert ungleich Null für die lokalen Umgebungen, in denen das Zielwort erscheint, in die Berechnung ein. Daraus ergibt sich eine Matrix für das Zielwort: es erscheint x-mal in Topik-1-Kontexten; es erscheint y-mal in Topik-2-Kontexten; es erscheint z-mal in Topik-3-Kontexten usw. Aus diesen Erscheinungshäufigkeiten ergibt sich dann ein prozentualer Anteil. Topik 1 hat einen Anteil von x Prozent an den Erscheinungskontexten des Wortes, Topik 2 hat einen Anteil von y Prozent an den Erscheinungskontexten des Wortes, Topik 3 hat einen Anteil von z Prozent an den Erscheinungskontexten des Wortes usw.

Betrachten wir eine Modellrechnung. Das Beispiel dient nur zur Veranschaulichung und geht der Einfachheit halber nur von fünf Topiks aus. Nehmen wir an, das Wort W_x erscheint in neun lokalen Umgebungen (hier als »Doc« abgekürzt), mit den folgenden Topiks mit Werten größer Null:

(1) - Doc1 (T1, T5) Wx (5 Topics)
- Doc2 (T2, T1)
- Doc3 (T1, T4)
- Doc4 (T1)

15 Siehe hierzu Roger Levy, T. Florian Jaeger: Speakers optimize information density.

- Doc5 (T1, T3)
- Doc6 (T3, T5)
- Doc7 (T2, T5)
- Doc8 (T2)
- Doc9 (T2, T1)

Diese Verteilung des Wortes in den lokalen Umgebungen des Gesamttexts übersetzt sich in folgende prozentuale Anteile der Topiks:

(2) - T1 6 0.375 Wx (5 Topics)
- T2 4 0.25
- T3 2 0.125
- T4 1 0.0625
- T5 3 0.1875

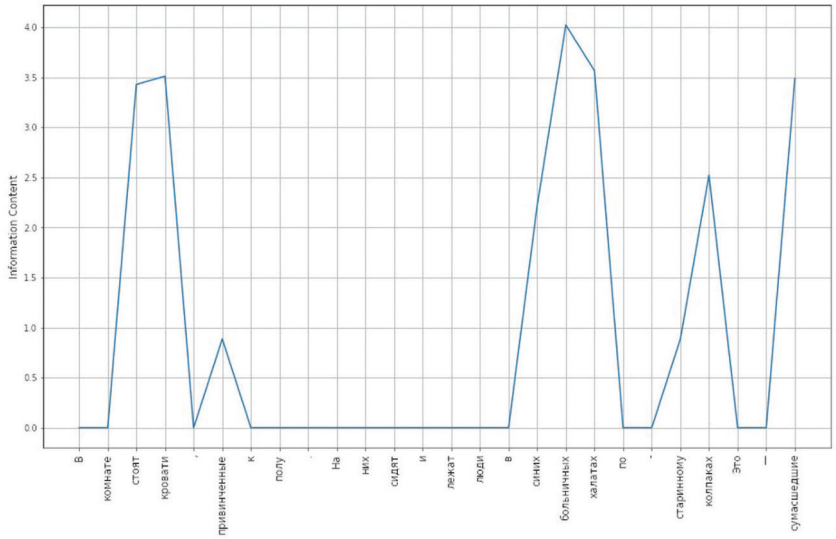
Aus diesen Werten lässt sich dann mit der Formel in (3) der Wert in (4) berechnen.

(3) $$\overline{SI}(\mathrm{w}) = -\frac{1}{n}\sum_{i=1}^{n} log_2 P(\mathrm{w} \mid \mathrm{t_i})$$

(4) TCM = -(0.2*log(0.375) + 0.2*log(0.25) + 0.2*log(0.125) + 0.2*log(0.0625) + 0.2*log(0.1875)) = 2,427552516

In unserem Beispielfall beträgt der Informationsgehalt des Wortes Wx im Text also 2,427552516.

Auf diese Weise kann für jedes Wort im Text dessen Informationsgehalt bestimmt werden. Wenn wir den Informationswert für jedes Wort in der Reihenfolge des Erscheinens der Wörter im Text serialisieren, bilden wir den Informationsfluss des Textes ab:



Etwas sehr salopp gesagt gewinnen wir auf diese Weise eine Art Überraschungs- oder Informationsprofil des Textes. Dargestellt wird, wie erwartet oder unerwartet die Wörter in der Erzählung vor dem Hintergrund der Topikstruktur des Textes erscheinen.

Der manipulierte Text

In der zweiten Phase des Projekts manipulieren wir den Gesamttext, indem wir ihn mit zusätzlichem Text anreichern. Der zusätzliche Text ist natürlich nicht willkürlich gewählt. Er soll das dem Originaltext implizite Hintergrundwissen explizieren.

Nochmal zurück zur Ausgangsidee. Jeder produzierte Text kommuniziert Botschaften, die sich aus zwei Bedeutungsquellen speisen. Zum einen aus den Bedeutungen, die die explizit genannten Wörter transportieren (sprachlich kodierte Bedeutungen). Zum anderen aus Bedeutungen, die die explizit genannten Wörter in ihrem jeweiligen Äußerungskontext zusätzlich triggern (kontextuell inferierte Bedeutungen). Das ist eine fundamentale Grundeigenschaft der menschlichen Sprache.

Čechov ist ein Autor, der die Technik, Bedeutungen kontextuell zu triggern, mit Meisterschaft beherrscht. Sein vielzitiertes Credo ›kratkost' – sestra talanta‹ (›Kürze ist die Schwester des Talents‹) lässt sich als programmatisch für seine Kunst verstehen. Das Talent besteht darin, den sprachlich kodierten Bedeutungsanteil der kommunizierten Botschaft zu minimieren und den kontextuell inferierten Bedeutungsanteil zu maximieren, ohne dabei den Kommunikationserfolg (die Vermittlung der Botschaft) zu gefährden.

Čechovsche Texte sollten sich demnach durch eine besonders große Kluft zwischen dem Gesagten und dem Gemeinten auszeichnen. Wenn man nur das explizit Gesagte beachtet, d. h., wenn man nur die sprachlich kodierten Bedeutungsanteile der Botschaft berücksichtigt, verpasst man Wesentliches. Im Grunde gilt das, wie gesagt, für jede sprachliche Äußerung. Nur stellen čechovsche Äußerungen in dieser Hinsicht einen Extremfall dar. Die Rede vom Subtext bei Čechov verweist auf die besondere Signifikanz des Impliziten in čechovschen Erzählungen (und Dramen).

Bei der oben in Abschnitt 4 beschriebenen Berechnung des Informationsflusses der Erzählung *Palata No. 6* wurde nur das explizite Wortmaterial der Erzählung berücksichtigt. Das heißt, dass dort nur sprachlich kodierte Bedeutungsanteile Eingang in die Rechnung gefunden haben. Das dort erzielte Ergebnis wollen wir nun mit jenem Informationsfluss vergleichen, den wir erhalten, wenn wir auch die kontextuell inferierten Bedeutungsanteile berücksichtigen.

Unsere Idee ist, dass wir das beschriebene Experiment wiederholen, nur dass wir den Text der Erzählung diesmal um Zusatztext anreichern. Der Zusatztext soll das implizite Wissen, das in der Erzählung steckt, explizieren. Wir wollen also, wenn man es so metaphorisieren will, den dunklen Text ins Licht rücken. Das ist natürlich ein gewaltiges und überambitioniertes Vorhaben, das höchstens annäherungsweise gelingen kann. Dennoch halten wir es für einen Versuch wert. Die Methode ist die folgende:

Inhaltswörter evozieren, wenn sie in einem Text gebraucht werden, Wissensframes.[16] Also suchen wir im Idealfall nach allen Wissensframes, die von den Inhaltswörtern der Erzählung evoziert werden. Frames stellen verallgemeinerte Formate für die mentale Repräsentation von Konzepten dar.[17] Als Ressource

16 Charles Fillmore: Frame Semantics and the Nature of Language. In: Annals of the NY Academy of Sciences: Conf. on the Origin and Development of Language and Speech 280 (1976). S. 20–32. Charles Fillmore: Frame semantics. In: The Linguistic Society of Korea (Hg.): Linguistics in the morning calm. Seoul 1982. S. 111–137.

17 Vgl. Dietrich Busse: Frame-Semantik. Ein Kompendium. Berlin u. a. 2012.

benötigen wir entsprechend eine Datenbank, die das mit Wörtern verknüpfte konzeptuelle Wissen verwaltet, d. h. eine Enzyklopädie. Als eine solche benutzen wir BabelNet.[18] Schnell wird klar, dass nicht alle Inhaltswörter der Erzählung als Lemmata in BabelNet vertreten sind. So müssen wir pauschal auf Adjektive und Verben verzichten. Nomina jedoch sind vertreten, und zu jedem Eintrag für ein Nomen hält BabelNet neben anderen Informationen eine Begriffsdefinition bereit. So lautet die Definition für das Nomen ›vrač‹ beispielsweise:

> (5) Vrač – vysšaja kvalifikacija specialista-lečebnika, pediatra, gigienista, ėpidemiologa, stomatologa; čelovek, ispol'zujuščij svoi navyki, znanija i opyt v profilaktike i lečenii zabolevanij, podderžanii normal'noj žiznedejatel'nosti organizma čeloveka.[19]

Ein von Matthias Irmer geschriebener Algorithmus extrahiert alle Definitionen der russischen BabelNet-Datenbank, die zu denjenigen Lemmata gehören, die mit Nomina in der Erzählung *Palata No. 6* korrespondieren. Die so gewonnenen Texte werden um Stoppwörter und Wiederholungen bereinigt (jedes Wort soll nur einmal vorkommen) und sodann im Originaltext an die Stelle ihres jeweiligen Triggerwortes eingefügt.

Auf diese Weise verändert sich der Originaltext natürlich gravierend. Erstens vergrößert sich der Textumfang erheblich. Zweitens wird der künstlerische Text durch die Einfügungen zerstört. Darüber hinaus sind viele Sätze syntaktisch nicht mehr wohlgeformt, und die Diskursstruktur des Textes ist inkohärent. Zur erbauenden Lektüre eignet sich der angereicherte Text also wohl kaum. Aber das soll er ja auch gar nicht. Für unsere Zwecke ist nur wichtig, dass er sich problemlos den Verfahren der Topikmodellierung mittels LDA und der Bestimmung des Informationsgehalts mittels TCM unterziehen lässt. Und das ist möglich, weil Topikmodellierungen ebenso wie das Topik-Kontext-Modell in ihren Kalkulationen Texte als ungeordnete Wortmengen (›bag of words‹) behandeln. Durch diese Eigenschaft der Modelle wird die Methode des Vergleichs eines reinen Texts und eines um automatisch extrahierte Wörter angereicherten Texts zu dem theoretischen Ansatz, den wir verfolgen.

---

18 Roberto Navigli, Simone Paolo Ponzetto: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. In: Artificial Intelligence Journal 193 (2012). Roberto Navigli, Simone Paolo Ponzetto: Multilingual WSD with just a few lines of code: the BabelNet API. In: Association for Computational Linguistics (Hg.): Proceedings of the ACL 2012 System Demonstrations (2012). S. 67–72.

19 URL: https://babelnet.org/synset?id=bn%3A00027976n&orig=vrac&lang=RU (9. Juni 2021).

Die beschriebene Anreicherung mit zusätzlichem Wortmaterial hat mutmaßlich Auswirkungen auf die Topikstruktur des Texts. Das sollte die Topikmodellierung ans Licht bringen. Die neue Topikstruktur wird dann, so unsere Hypothese, auch die Informationswerte der Wörter, die wir auf der ersten Etappe gemessen haben, verändern.

Ich fasse die Projektidee noch einmal zusammen: Zunächst bestimmen wir den Informationsgehalt der Wörter der Erzählung auf der Basis des Originaltexts und erstellen daraus ein Informationsprofil. Dann stellen wir diesem ein zweites Informationsprofil gegenüber. Letzteres erhalten wir, indem wir den Informationsgehalt der Wörter noch einmal messen, diesmal auf der Basis des wie oben beschrieben manipulierten Texts. Die Differenz zwischen dem ersten Informationsprofil und dem zweiten Informationsprofil ist, so die Annahme, ein Effekt des Subtexts der Erzählung *Palata No. 6*.

Blick ins Tal

Hier endet mein Bericht. Er handelt in vielerlei Hinsicht von einer Gratwanderung, auf der wir uns derzeit befinden. Die Kammlinie trennt die Linguistik auf der einen Seite von der Literaturwissenschaft auf der anderen Seite. Das habe ich oben beschrieben. Eine Gratwanderung ist das Projekt aber auch unter einem anderen Gesichtspunkt. Sie wird vermutlich alle Forscherinnen und Forscher, die sich um ein theoretisches Verständnis sprachlicher Bedeutungen bemühen, in zwei Lager spalten – ob sie nun linguistischer oder literaturwissenschaftlicher Provenienz sind.

Die einen werden dem geschilderten Forschungsansatz mit größter Skepsis begegnen. Was soll bei einer digitalen Textanalyse des Subtexts schon anderes herauskommen als eine Variation der Antwort auf die ultimative Frage (die bekanntlich »42« lautet)? Bedeutungen, insbesondere Interpretationen, sind schließlich genuine Leistungen des menschlichen Geistes. Diesen quantifizieren zu wollen kann ja nur ein Holzweg sein!

Auf der anderen Seite des Grats werden sich die versammeln, die die neuen Methoden der Digital Humanities verheißungsvoll finden. Natürlich, im aktuellen Entwicklungsstadium auf diesem Gebiet sind noch keine tiefen Erkenntnisse zu erwarten. Aber die Entwicklungen sind rasant, und wer weiß, vielleicht werden wir in gar nicht allzu ferner Zukunft nicht nur präzise nachweisen können, *dass* eine semantische Kluft zwischen der Bedeutung eines Texts ohne und mit Subtext besteht (was wir ja schon wissen). Vielleicht werden wir zunächst grob, dann immer feiner aufzeigen können, worin die textspezifische Kluft besteht.