Datengeleitete Kategorienbildung in den Digital Humanities: Paraphrasen aus korpus- und computerlinguistischer Perspektive

Joachim Scharloth / Franz Keilholz / Simon Meier-Vieracker / Xiaozhou Yu / Roman Dorniok

Abstract Der Begriff der Paraphrase wird in unterschiedlichen Disziplinen mit jeweils eigenem Erkenntnisinteresse verwendet und mit je unterschiedlichen Akzenten als ein spezieller Fall der Intertextualität definiert. Angesichts der Vielfalt der definitorischen Ansätze und der damit einhergehenden Unschärfe des Begriffs stellt sich die Frage, inwieweit maschinelle Ansätze zur Identifizierung von Paraphrasen überhaupt erfolgreich sein können.

Algorithmen in den Digital Humanities können als Modellierungen von Konstrukten aufgefasst werden. Die Anpassung der Parameter zur Steuerung der Algorithmen ist deshalb kein technischer, sondern ein interpretativer Vorgang, bei dem in einer Schleife aus Parameteranpassung und Interpretation des Outputs eine wechselseitige Präzisierung von Parametern und Kategorienbildung erfolgt.

Unser Beitrag illustriert am Beispiel der n-Gramm-basierten Paraphrasensuche in griechisch- und deutschsprachigen Korpora, wie eine Variation der Parameter zur Modellierung unterschiedlicher Paraphrasenbegriffe führt und inwiefern die Interpretation des Algorithmus zu einer Präzisierung der jeweiligen Paraphrasenbegriffe beitragen kann.

Keywords Paraphrase, Korpuslinguistik, Methodologie der Digital Humanities, n-Gramm-basierte Paraphrasensuche

Einleitung

In Douglas Adams' Science-Fiction-Klassiker *The Hitchhiker's Guide to the Galaxy* wird neben vielen anderen skurrilen Episoden aus den Weiten des Weltalls in Band 3 der Romanserie[1] auch die Geschichte des Dichters Lallafa erzählt, die im Hinblick auf das Problem der Paraphrase einige interessante Anknüpfungspunkte enthält. Der Dichter Lallafa lebte demnach wie ein Einsiedler in den Wäldern der Langen Länder von Effa und schrieb seine Gedichte auf die getrockneten Blätter der Hafra-Pflanze. Er schrieb über die Schönheiten der Natur und über seine unerfüllte Liebe zu einem Mädchen, das ihn verlassen hatte. So einsam wie er lebte, so starb er auch; und so kam es, dass erst lange nach seinem Ableben seine Gedichte gefunden und publiziert wurden. Sie wurden vom Publikum so positiv aufgenommen, dass er postum zu großem Ruhm kam und seine Gedichte über viele Jahrhunderte als die besten in der ganzen Galaxis galten.

Nach der Erfindung des Zeitreisens kamen Hersteller von Korrektur-Fluid auf den Gedanken, dass die Gedichte Lallafas womöglich hätten noch besser sein können, hätte er ein Korrektur-Fluid zur Verfügung gehabt. Sie reisten zu ihm in die Vergangenheit, um ihn davon zu überzeugen, sich zu Werbezwecken positiv über das Korrektur-Fluid zu äußern – selbstverständlich gegen Bezahlung. So kam es, dass Lallafa es zu außerordentlichem Reichtum brachte. Häufig pendelte er auch in die Zukunft, um in Talkshows aufzutreten. Der Reichtum, den ihm die Werbeverträge einbrachten, wirkte sich freilich auf Lallafas Leben aus: Das Mädchen, das er liebte, kam gar nicht mehr auf die Idee, ihn zu verlassen, und Lallafa entschied sich, nicht in den Wäldern zu wohnen, sondern leistete sich eine schicke Wohnung in der Stadt. All das führte dazu, dass er jene Gedichte, die ihn berühmt machen würden, gar nicht erst zu schreiben begann. Dies wollten freilich die Hersteller des Korrektur-Fluids nicht hinnehmen und gaben ihm ein Exemplar einer in der Zukunft erschienenen Ausgabe seiner Gedichte und einige Hafrablätter, damit er die Gedichte abschriebe, freilich nicht ohne Änderungen anzubringen, die den Gebrauch von Korrektur-Fluid nötig machten.

So fertigte Lallafa Überarbeitungen seiner eigenen Gedichte an, die zwar als seine ursprünglichen Gedichte vertrieben wurden, die jedoch, um als Werbung für das Korrektur-Fluid gelten zu können, als Überarbeitungen seiner in der Zukunft erschienenen ursprünglichen Gedichte erkennbar sein mussten. Bei seinen Leserinnen und Lesern sorgten die Gedichte für ein geteiltes Echo: Für einige waren seine Gedichte wertlos geworden, andere behaupteten, sie seien genauso, wie sie immer waren.

1 Adams (1984) Kapitel 17.

Das durch das Zeitreisen erzeugte Paradox, das die Autorinnen und Autoren in ihrer Paraphrase der Romanpassage nicht ganz ohne Korrektur-Fluid dargestellt haben, verweist auf einige Fragen, mit der sich auch die Forschung zu Paraphrasen beschäftigt: Mit der gespaltenen Leserschaft der überarbeiteten Gedichte kann man sich fragen, inwiefern eine Paraphrase etwas Neues ist – oder lediglich das Gleiche in einem anderen Gewand. Oder noch radikaler: Wann etwas überhaupt eine Paraphrase ist und wann etwas anderes? Mit den Herstellern des Korrektur-Fluids ließe sich fragen, ob eine Paraphrase stets eine materielle Beziehung zum Prätext hat, also quasi als dessen Bearbeitung sichtbar sein muss – oder ob eine Textpassage auch ohne einen Bezug an der Textoberfläche zu einem vermeintlichen Prätext als Paraphrase gelten kann. Lallafas Paraphrase seiner eigenen Gedichte, die er angesichts seines Reichtums nicht mehr geschrieben hätte, wirft zudem die Frage auf, wer in der Paraphrase eigentlich spricht. Ist es der ursprüngliche Autor, der im Falle Lallafas wegen des Zeitparadoxes freilich gar nicht mehr als einsiedlerisch lebender Autor existiert – oder sind es kulturell geprägte Muster, denen frühere Texte lediglich Ausdruck verliehen haben und die nun auch in der Paraphrase reproduziert werden? Das Paradox wirft damit auch Fragen nach der Rolle des Autors auf, der entweder origineller und intentionaler Schöpfer ist – oder lediglich Sprachrohr für bereits vorhandene Aussagen in einem sprachlichen Möglichkeitsuniversum.

Wie die folgenden Ausführungen zur theoretischen Reflexion des Paraphrasenbegriffs zeigen werden, identifiziert die Geschichte des Dichters Lallafa die wichtigsten Problemfelder des Intertextualitätsdiskurses.

Die Paraphrase als Sonderfall der Intertextualität

Von Paraphrasen zu sprechen impliziert, eine bestimmte Form der Intertextualität anzunehmen. Dabei ist das Phänomen der Intertextualität in unterschiedlichen Disziplinen und Denkschulen auf sehr unterschiedliche, ja teils einander ausschließende Weisen theoretisiert worden. Die Vielzahl der Ansätze zur Bestimmung des Phänomens der Intertextualität lässt sich grob in deskriptive und ontologische Theorien gliedern.

Deskriptive Intertextualitätstheorien nehmen Beziehungen zwischen Texten in den Blick, soweit sich diese als intentionale Referenzen eines Autors auf einen anderen Text deuten lassen. Der Literaturwissenschaftler Wolfgang Stierle, der vehement für eine „deskriptive, auf das je einzelne Verhältnis bezogene Kategorie“[2] der Intertextualität plädierte, fasste Zitat, Übersetzung und Anspielung, Parodie

2 Stierle (1984) 147.

und Travestie, sowie den Kommentar, die Interpretation und die Kritik[3] als typische Formen deskriptiver Intertextualität. Die Paraphrase kann man hier in den Praktiken der Übersetzung und Anspielung oder als Element der Schreibweisen der Parodie und Travestie sowie als funktionaler Bestandteil der Gattungen des Kommentars, der Interpretation oder der Kritik verorten. Je nach dem, welchen Einfluss der Text-Text-Bezug auf die Gestaltung des Prätextes hat, differenzieren deskriptive Intertextualitätstheorien in eine Produktions- und Rezeptionsintertextualität. Während bei letzterer zwar die Sinnkonstitution latent beeinflusst, die Oberfläche jedoch nicht maßgeblich mitprägt wird, wird bei der ersteren die Textoberfläche von der intentionalen Bezugnahme mitstrukturiert.[4]

Ontologische Intertextualitätstheorien wurzeln in Michail Bachtins Theorie der Dialogizität. Folgt man dem russischen Literaturwissenschaftler, so tragen Wörter nicht nur eine Bedeutung, die durch strukturelle Differenzen zu anderen Zeichen des Sprachsystems bestimmt ist. Vielmehr haben sich in sie „Gebrauchswerte"[5] eingeschrieben, Bedeutungsdimensionen, die sich aus den historischen Verwendungsweisen in konkreten Kontexten ergeben. Bachtin schreibt:

> Es [das Wort] vermag sich nicht restlos aus der Gewalt jener Kontexte zu lösen, in die es einst einging. Jedes Mitglied eines Sprechkollektivs findet das Wort nicht als ein neutrales Wort der Sprache vor, das von fremden Bestrebungen und Bewertungen frei ist, dem keine fremden Stimmen innewohnen. Nein, es empfängt das Wort von einer fremden Stimme, angefüllt mit dieser fremden Stimme. In seinen Kontext kommt das Wort aus einem anderen Kontext, durchwirkt von fremden Sinngebungen. Sein eigener Gedanke findet das Wort bereits besiedelt.[6]

Jedes eigene Sprechen ist also unhintergehbar ein Dialog mit früheren Verwendungen des sprachlichen Materials und den durch sie mitgeprägten Verstehenshorizonten der Adressatinnen und Adressaten.

Aus diesem Gedanken Bachtins entwickelte die französische Literaturwissenschaftlerin Julia Kristeva ihren Begriff der Intertextualität, der die literaturwissenschaftliche Debatte maßgeblich geprägt hat. So wie bei Bachtin Sprache bereits sozial geprägt ist, so ist Intertextualität die Seinsform von Texten schlechthin und nicht nur die intendierte Bezugnahme auf einen Prätext: „[…] jeder Text baut sich als Mosaik von Zitaten auf, jeder Text ist Absorption und Transformation eines anderen Textes."[7]

---

3 Vgl. ebd. 147.
4 Vgl. Lachmann (1984) 134.
5 Feilke (2000) 78.
6 Bachtin (1990) 130.
7 Kristeva (1978) 391.

Ein so gefasster Intertextualitätsbegriff fügt sich dem Programm der poststrukturalen Literaturwissenschaft insofern ein, als er die Vorstellung von Subjektautonomie und auktorialer Intentionalität als Quelle der Bedeutung eines Textes obsolet macht. In Roland Barthes' berühmtem Essay *Der Tod des Autors* (1968) heißt es entsprechend: „Ein Text ist aus vielfältigen Schriften zusammengesetzt, die verschiedenen Kulturen entstammen und miteinander in Dialog treten, sich parodieren, einander in Frage stellen. Es gibt aber einen Ort, an dem diese Vielfalt zusammentrifft, und dieser Ort ist nicht der Autor (wie man bislang gesagt hat), sondern der Leser."[8] Der Leser freilich ist nicht mehr als ein

> Raum, in dem sich alle Zitate, aus denen sich Schrift zusammensetzt, einschreiben, ohne dass ein einziges verloren ginge. Die Einheit eines Textes liegt nicht in seinem Ursprung, sondern in seinem Zielpunkt – wobei dieser Zielpunkt nicht mehr länger als eine Person verstanden werden kann. Der Leser ist ein Mensch ohne Geschichte, ohne Biographie, ohne Psychologie. Er ist nur der Jemand, der in einem einzigen Feld alle Spuren vereinigt, aus denen sich das Geschriebene zusammensetzt.[9]

Aus der Perspektive der ontologischen Intertextualitätstheorie sind Paraphrasen universelle Eigenschaften von Texten, weil Text und Intertextualität letztlich in eins gesetzt werden. Texte verweisen in einem *regressus ad infinitum* immer wieder auf andere Texte und letztlich auf alle anderen Texte des Textuniversums. Ein solch totaler Paraphrasenbegriff ist zwar in theoretischer Hinsicht reizvoll, insofern er eine Verschiebung des Blickwinkels auf Text, Autor und Leser bedingt, jedoch mangels Differenzierung des Phänomenbereichs kaum für spezifische Forschungsfragen geeignet.

Eine Neuordnung der zwischen geistesgeschichtlich-deskriptiven und literaturtheoretisch-ontologischen Ansätzen gefangenen Intertextualitätstheorie schlug der Literaturwissenschaftler Gérard Genette Anfang der 1980er vor,[10] indem er fünf Grundtypen transtextueller Beziehungen identifizierte.

Den Begriff der Intertextualität schränkte er dabei auf einen bestimmten Typ transtextueller Beziehung ein: „Ich definiere sie wahrscheinlich restriktiver als Beziehung der Kopräsenz zweier oder mehrerer Texte, d. h. in den meisten Fällen, eidetisch gesprochen, als effektive Präsenz eines Textes in einem anderen".[11] Erscheinungsformen der Intertextualität sind das Zitat als wörtliche, als solche deklarierte Übernahme, das Plagiat als wörtliche, nicht-deklarierte Übernahme und

8 Barthes (2000) 190.
9 Barthes (2000) 192.
10 Vgl. Genette (1993) [frz. Erstauflage 1982].
11 Genette (1993) 10.

die Anspielung als fragmentarische, nicht-deklarierte Entlehnung, die sich nur jenen Leserinnen und Lesern erschließt, wenn der Text, auf den die Anspielung Bezug nimmt, ihm bzw. ihr bekannt ist.

Als zweiten Typus transtextueller Relationen identifiziert Genettes Taxonomie die Paratextualität, mit der er die Ausstattung eines Textes mit Co-Texten wie Umschlagtext, Titel, Untertitel, Vorwort, Einleitung, Motti, Marginalien, Fußnoten, Anmerkungen, Nachwort fasst. Metatextualität als der dritte Typus transtextueller Relationen bezeichnet die „Beziehung zwischen einem Text und einem anderen, der sich mit ihm auseinandersetzt, ohne ihn unbedingt zu zitieren".[12] Genettes Fokus liegt auf der Hypertextualität als viertem Typus transtextueller Beziehungen, bei dem ein Text einen zeitlich vor ihm erstellten Text zur Folie macht, etwa in Form einer Parodie, eines Pastiches, einer Adaption,[13] ohne dabei Kommentar im Sinne der Metatextualität zu sein. Den fünften Typ transtextueller Relationen bestimmt Genette als Architextualität, die die Zugehörigkeit eines Textes zu einer bestimmten Textsorte oder einem Genre bezeichnet.[14]

Mit dieser Taxonomie ordnet Genette das Feld transtextueller Bezüge neu und schränkt insbesondere die Extension des Intertextualitätsbegriffs ein, die in der Präsenz eines Textes in einem anderen durch Zitat, Plagiat oder Anspielung besteht. Damit besteht eine große Ähnlichkeit zur Einzeltextreferenz als Spielart der Intertextualität, die von Manfred Pfister[15] und Ulrich Broich[16] zusammen mit der Systemreferenz als Haupttypen der Intertextualität bestimmt wurde. Demnach liegt Einzeltextreferenz vor, wenn „ein Text [sich] auf einen bestimmten, individuellen Prätext [bezieht]".[17] Unter den traditionellen Termini für diese Form der Intertextualität findet sich bei Broich auch die Paraphrase: „Zitat, Motto, Cento, Übersetzung, Bearbeitung, *imitation* (im klassizistischen Sinn), Paraphrase, Resümee, Kontrafaktor und viele andere mehr."[18] Die Einzeltextreferenz als „bewußte, intendierte und markierte Intertextualität"[19] rechnet Pfister zum Kernbereich der Intertextualität. Der Gegenbegriff zur Einzeltextreferenz ist die Systemreferenz. Intertextueller Bezugsrahmen ist hier „nicht mehr ein individueller Prätext, sondern wird von Textkollektiva gebildet oder genauer von den hinter ihm stehenden und sie strukturierenden Systemen".[20]

12 Genette (1993) 13.
13 Vgl. Aczel (2004) 112.
14 Vgl. Genette (1993) 13.
15 Vgl. Pfister (1985a) und (1985b).
16 Vgl. Broich (1985a) und (1985b).
17 Broich (1985b) 52.
18 Broich (1985b) 49; Hervorhebung im Original.
19 Broich (1985b) 48.
20 Pfister (1985b) 53.

Auch in der Sprachwissenschaft äußerten insbesondere die anwendungsbezogenen Bereiche wie die Textlinguistik und die Korpuslinguistik Vorbehalte gegen die ontologische Intertextualitätstheorie und diskutierten eine operationalisierbare Eingrenzung des Intertextualitätsbegriffs. So forderte Wolfgang Heinemann, den Begriff der Intertextualität einzuschränken auf die Wechselbeziehungen zwischen konkreten Texten und die grundsätzliche Textsortengeprägtheit aller Texte, während er für das allgemeine Phänomen der auch impliziten Text-Text-Beziehungen den Terminus der Textreferenz vorschlug.[21] Mit Holthuis unterscheidet er textoberflächenstrukturelle und texttiefenstrukturelle Referenzen. Erstere richten sich „(dominant) auf die *linearisierte Version* des Bezugstextes" und umfassen beispielsweise das wörtliche Zitat, letztere hingegen richten sich „(dominant) auf die *nichtlinearisierten Eigenschaften* des Bezugstextes"[22] und werden beispielsweise in der Anspielung, der Paraphrase, der Übersetzung und der Bearbeitung realisiert.

Kathrin Steyer bestimmte den „Kernbereich sprachwissenschaftlichen Interesses" im Feld der Intertextualitätsforschung als „Referenzen auf Versprachlichtes" im Sinne von Reformulierungen. Als empirische Wissenschaft müsse die Linguistik sprachproduktbezogene Intertextualität in das Zentrum ihrer Betrachtung stellen. Eine analytische Beschreibung müsse „auch ohne Instrumentarien der Psychologie, kognitivistischer Theorien oder gar der Literaturwissenschaft geleistet werden" können.[23] Entsprechend müsse nicht notwendig der Nachweis intentionaler Reformulierungen und direkter Textbezüge im Zentrum des Interesses stehen, sondern die Distribution textueller Muster in größeren Korpora und ihre diachrone Entwicklung. Der Musterbegriff impliziert dabei, dass Prätext und Paraphrase auf dieselbe abstrakte Repräsentation zurückgeführt werden können und diese abstrakte Repräsentation als das Vehikel ihrer wechselseitigen Transformation gedacht werden kann.

Lässt man die unterschiedlichen Ansätze zur Bestimmung von Intertextualität und zur Verortung der Paraphrase in einer umfassenden Theorie der Intertextualität Revue passieren, so zeigen sich vier Tendenzen: Erstens ist in der Forschung das Bemühen um eine operationalisierbare Differenzierung des terminologischen Apparats festzustellen. Deskriptive und ontologische Traditionen bleiben aber insofern prägend, als sich zweitens das Feld der Intertextualitätsforschung zwischen den Polen der Einzeltextorientierung einerseits und der Systemorientierung andererseits aufspannt. Drittens wird die Paraphrase von allen Theoretikern als Phänomen der Einzeltextreferenz bestimmt. Viertens ist auch im Hinblick auf den Begriff der Paraphrase eine Polysemie zu konstatieren. Während eine Tradition die Intentionalität der Referenz auf einen einzelnen Prätext zu einem wesentlichen

21 Vgl. Heinemann (1997) 35.
22 Holthuis (1993) 91.
23 Steyer (1997) 86.

Bestandteil der Paraphrasendefinition erhebt (Pfister / Broich), genügt der anderen die Ähnlichkeit des sprachlichen Produkts, um von einer Paraphrase sprechen zu können (Steyer). Entsprechend kann sich das mit der Paraphrasenanalyse verbundene Erkenntnisinteresse einerseits auf die „Textarchäologie“[24] richten, andererseits aber auch auf die Begriffs- und Diskursgeschichte und die Geschichte von Topoi, d. h. eine Geschichte von Themen- und Argumentationen. Je nach theoretischer Verortung richtet sich eine diskurs- und toposorientierte Paraphrasenforschung damit auf ein den Prätext (und mit ihm die Paraphrase) strukturierendes System von Topoi, Aussagen, Motiven oder Codes, das allerdings nicht als bloßes Postulat universeller Intertextualität gedacht wird, sondern in den sprachlichen Oberflächen- oder Tiefenstrukturen nachweisbar sein muss.

Im Folgenden soll diskutiert werden, inwiefern das Kriterium der Transformierbarkeit geeignet ist, die weiter bestehende Polysemie des Paraphrasenbegriffs einzuschränken.

Sprachliche Transformationsdimensionen

Ganz gleich, ob man Paraphrasen als intentionale und zumindest implizit markierte Reformulierungen oder als rein sprachliche Homologien konzeptualisiert, so bleibt doch der Bezug auf die Transformierbarkeit von Teilen des Prätextes in die Paraphrase ein wichtiger Bestandteil der Paraphrasendefinition.

In der linguistischen Forschung gibt es insbesondere in der Computerlinguistik und dem *Natural Language Processing* unterschiedliche Ansätze zur Klassifizierung dieser Transformationsvorgänge. Etwa hat Dras[25] differenzierte Vorschläge zur syntaktischen Klassifikation von Paraphrasen vorgelegt und dabei 54 unterschiedliche Typen identifiziert. Bhagat[26] macht die lexikalischen Veränderungen zwischen Prätext und Paraphrase zum Hauptklassifikationskriterium. Den anspruchsvollsten Versuch zur Klassifikation hat Fujita[27] unternommen, der grundsätzlich zwischen lexikalischen und strukturellen Paraphrasen unterscheidet.

Entsprechend unseres Verständnisses von einer Paraphrase als durch die Transformierbarkeit des Prätexts in den Paraphrasentext konstituiertes Intertextualitätsphänomen, halten wir eine an sprachlichen Formen auf unterschiedlichen Sprachrängen orientierte Klassifikation für besonders operabel. Lose angelehnt an

24 Pfister (1985a) 23.
25 Vgl. Dras (1999) 59–75.
26 Vgl. Baghat (2009) 30–45.
27 Vgl. Fujita (2005) 11–19.

die Zusammenstellung bei Vila et al.,[28] die auf die genannten Klassifikationsversuche referiert, können wir unterschiedliche Transformationsdimensionen unterscheiden, die im Folgenden exemplarisch vorgestellt werden sollen.

Auf der Ebene der *Morpholexik* lassen sich einerseits morphologische, andererseits lexikalische Transformationen unterscheiden. *Morphologische Transformationen* umfassen beispielsweise Unterschiede im Numerus („Ansturm der Feinde“ > „Ansturm des Feindes“) oder in der Kasusmarkierung („In gleichem Maße“ > „In gleichem Maß“), aber auch derivationale Veränderungen („Die Vorgänge sind ohne Übertreibung zu behandeln“ > „Die Vorgänge sind ohne Übertreiben zu behandeln“). *Lexikalische Transformationen* können unter anderem die Verwendung von Quasi-Synonymen („die wutentbrannten Äußerungen der britischen Presse“ > „die wütenden Äußerungen der britischen Presse“) umfassen, wobei zwischen der Substitution von gleicher („hervorragende Leistungen“ > „exzellente Leistungen“) und entgegengesetzter Polarität („nicht-militärische Ziele“ > „zivile Ziele“) unterschieden werden kann. Ebenso können in einer Paraphrase Hyperonyme durch Hyponyme und Hyponyme durch Hyperonyme ersetzt werden („neuartiger Sprengkörper“ <> „neuartige Waffe“). Auch Komposita können durch Umschreibungen ersetzt werden und umgekehrt („Kriegsziele“ <> „Ziele des Kriegs“).

Durch bestimmte Transformationen auf der *morphosyntaktischen Ebene* kann in der Paraphrase auch die *Modalität* des Prätextes verändert werden. Etwa können direkte in indirekte Fragen umformuliert werden („Kann der Sieg auf diese Weise errungen werden?“ <> „Es stellt sich die Frage, ob der Sieg auf diese Weise errungen werden kann.“) und umgekehrt. Ebenso können unmarkierte Aussagen durch Modalverben in Vermutungsbedeutung in eine epistemische Modalität („Die Invasion hat ihren Anfang genommen.“ > „Die Invasion muss ihren Anfang genommen haben.“) oder durch Modalverben mit der Bedeutung einer fremden Behauptung oder den Gebrauch indirekter Rede in einen Ausdruck von Evidentialität („Die Invasion hat ihren Anfang genommen.“ <> „Die Invasion soll ihren Anfang genommen haben.“ <> „Das OKW berichtet, die Invasion habe ihren Anfang genommen“) umgeformt werden und umgekehrt. Ebenso kann es in der Paraphrase zu einer Transformation der *Diathese* kommen, im Deutschen vorwiegend vom Aktiv zum Passiv und umgekehrt („Deutsche Verbände haben die Bombardierung Londons fortgesetzt.“ <> „Die Bombardierung Londons wurde fortgesetzt.“).

Auf der Ebene der *Syntax* lassen sich grob Veränderungen in der Wortstellung und Ersetzungen von Satzgliedern und Gliedsätzen unterscheiden. *Veränderungen in der Wortstellung* kann eine Verschiebung des Fokus zur Folge haben („Der Feind will uns in diesem Jahr militärisch überrennen“ > „In diesem Jahr will der Feind uns militärisch überrennen“). Die *Substitution von Satzgliedern und Gliedsätzen* kann

28 Vila et al. (2014) 211–213.

beispielsweise die Ersetzung von Präpositionalgruppen durch Adverbien („wird mit aller Schärfe widersprochen“ > „wird schärfstens widersprochen“), Transformationen von subordinierenden in koordinierende Nebensätze und umgekehrt („obwohl unsere Feinde in der Überzahl sind, schlagen sich unsere Truppen …“ <> „Unsere Feinde sind in der Überzahl, trotzdem schlagen sich unsere Truppen …“) oder die Ersetzung von Relativsätzen durch Partizipialkonstruktionen und umgekehrt („Die erbitterten Kämpfe, die höchste Anforderungen an die deutsche Verteidigungskraft stellen“ <> „Die höchsten Anforderungen an die deutsche Verteidigungskraft stellenden Kämpfe“) umfassen. Selbstverständlich können in der Paraphrase auch einzelne Sätze des Prätextes miteinander verbunden oder längere Sätze des Prätextes in mehrere Einzelsätze geteilt werden („Mit der Bekanntgabe der Namen ist das Thema abgeschlossen. Eine weitere Erörterung erübrigt sich daher.“ <> „Mit der Bekanntgabe der Namen ist das Thema abgeschlossen, weswegen sich eine weitere Erörterung erübrigt.“). Satzglieder können weggelassen oder ergänzt werden.

Daneben können sich in der Paraphrase auch Veränderungen in der *graphematischen Realisierung* von Lexemen und Kognaten („&“ <> „und“, „%“ <> „Prozent“) finden sowie Varianten in der *Orthographie* („zurückgetreten“ <> „zurück getreten“).

Wie bei fast jeder sprachwissenschaftlichen Taxonomie ist eine saubere Trennung formaler und funktional-inhaltlicher Dimensionen kaum möglich, so dass es durchaus zu Überlappungen zwischen den Dimensionen kommen kann; zudem kommen bei der Paraphrasierung häufig mehrere Transformationsstrategien parallel zum Einsatz.

Nimmt man die Transformationsdimensionen zum Ausgangspunkt einer operationalen Bestimmung des Paraphrasenbegriffs, so wird deutlich, dass auch in dieser Perspektive Paraphrasen ein breites Spektrum sprachlicher Phänomene abdecken können, das von der Textübernahme mit minimalen Änderungen bis hin zur bloß semantischen Ähnlichkeitsbeziehung zum Prätext reicht. Für letzteres sind insbesondere die potenziellen Transformationen im Bereich der Lexikosemantik verantwortlich: Transformationen, die Hyperonym-, Hyponym- oder Kohyponomiebeziehungen zwischen lexikalischen Einheiten von Prätext und Paraphrase einschließen, können in Verbindung mit anderen Transformationen zu Paraphrasen führen, bei denen die Einzeltextreferenz (trotz sprachlicher Homologie und wechselseitiger Transformierbarkeit zwischen Prätext und Paraphrase) in den Hintergrund tritt und sich die Systemreferenz als interpretativer Bezugsrahmen aufdrängt. So lässt sich beispielsweise der Satz „Der Reichskanzler steht stets zu seiner Aussage“ als Paraphrase eines Prätextes auffassen, der den Ausdruck „Ein Mann, ein Wort“ enthält, denn beide Sätze lassen sich über Hyperonymbeziehungen zwischen den Nomen wechselseitig ineinander transformieren. Selbst wenn wir in diesem Fall einen intentionalen Einzeltextbezug annehmen würden,

wäre doch die Interpretation, dass die Referenz hier weniger der Aussage des Prätextes gilt, sondern vielmehr dem dahinter stehenden Topos von der vermeintlichen Worttreue des männlichen Geschlechts, also mithin ein Systembezug, die näherliegende Deutung.

Dagegen ließe sich wiederum einwenden, dass solche Hyperonymbeziehungen wie die zwischen „Mann" und „Reichskanzler" zu abstrakt sind, als dass sie geeignet wären, als Indikatoren für Einzeltextreferenzen gelten zu können. Allerdings wäre dies eine willkürliche Grenzziehung, die sich kaum auf sprachwissenschaftliche Kriterien zurückführen ließe, immerhin scheint die Hyperonomiebeziehung zwischen „Aussage" und „Wort" weit weniger problematisch.

Das Kriterium der Transformierbarkeit ermöglicht damit keine Begrenzung der Polysemie des Paraphrasenbegriffs. Was dies für die maschinelle Paraphrasenidentifizierung bedeutet, soll in den folgenden Abschnitten diskutiert werden.

Maschinelle Paraphrasensuche: Methodologie und Operationalisierung

Konstrukt und algorithmische Operationalisierung

Die maschinelle Paraphrasensuche sieht sich folglich mit dem Problem einer doppelten Unschärfe konfrontiert: Einerseits ist der Paraphrasenbegriff trotz seiner Beschränkung auf die Kriterien der sprachproduktbezogenen Homologie und der wechselseitigen Transformierbarkeit unscharf im Hinblick auf seine Offenheit für Systemreferenzen; andererseits ist das formale Kriterium der sprachlichen Transformierbarkeit bzw. der Rückführung von Prätext und Paraphrase auf ein gemeinsames sprachliches Muster unscharf im Hinblick darauf, bis hin zu welcher Abstraktionsstufe Transformationsprozesse als paraphrasenwertig gelten können.

Diese Unschärfe ist freilich in den Geistes- und Kulturwissenschaften nicht ungewöhnlich – sie ist vielmehr konstitutives und produktives Merkmal ihres Forschungsprozesses. Seit Quines Dekonstruktion der Unterscheidung von analytischen und synthetischen Wahrheiten[29] und mit dem damit einsetzenden *linguistic turn* hat sich die Einsicht durchgesetzt, dass konzeptuelle und historisch-empirische Erkenntnisweise stets zusammenfließen, dass Erfahrung und die unhintergehbare sprachliche Bedingtheit unseres Denkens nicht zu trennen sind und dass daher Wahrheiten *de dicto* und Wahrheiten *de re* nicht ohne Weiteres unterschieden

29 Vgl. Quine (1951).

werden können. Entsprechend gehen im Forschungsprozess Kategorienbildung und Auseinandersetzung mit dem empirischen Forschungsgegenstand notwendig Hand in Hand.[30]

Dies gilt auch für die Digital Humanities. Zwar können Algorithmen in den Digital Humanities als Modellierungen von Konstrukten wie dem der Paraphrase aufgefasst werden. Die schrittweise Entwicklung der Algorithmen selbst und die Anpassung der Parameter zu ihrer Steuerung ist jedoch kein technischer, sondern ein interpretativer Vorgang, bei dem in einer Schleife aus Parameteranpassung und Interpretation des Outputs eine wechselseitige Präzisierung von Parametern und Kategorienbildung erfolgt.

Die maschinelle Paraphrasenidentifizierung für geistes- und kulturwissenschaftliche Fragestellungen sucht also zunächst nach einem algorithmischen *Modell* der Paraphrase. Modelle sind nach Stachowiak durch ihren Abbildcharakter („Modelle sind stets Modelle von etwas, nämlich Abbildungen, Repräsentationen natürlicher oder künstlicher Originale"), die Notwendigkeit der Verkürzung („Modelle erfassen […] nicht alle Attribute des durch sie repräsentierten Originals, sondern nur solche, die den jeweiligen Modellerschaffern und / oder Modellbenutzern relevant scheinen"[31]) und dem ihnen innewohnenden Pragmatismus (Modelle erfüllen Ersetzungsfunktion für bestimmte Subjekte, in einer bestimmten Zeit und eingeschränkt auf bestimmte Operationen) charakterisiert.

Modelle für die geistes- und kulturwissenschaftliche Textanalyse müssen die Besonderheiten ihrer Gegenstände in allen drei Bereichen berücksichtigen. Dies bedeutet, der Tatsache Rechnung zu tragen, dass der *Abbildcharakter* des Modells sich in den Geistes- und Kulturwissenschaften auf interpretative Kategorien (wie das Konzept der Paraphrase) bezieht, Modelle also Interpretationen von Interpretationen (im Sinne von Geertz[32]) sind, mithin ein konstruktorientierter Modellbegriff zugrunde gelegt werden muss. Die Forschungslogik von Definition, algorithmischer Operationalisierung und Gütemessung mittels *Precision* und *Recall* kann damit bestenfalls zur Validierung der zahlreichen Zwischenschritte des zirkulären Forschungsprozesses dienen, ist aber im Ganzen den Gegenständen geistes- und kulturwissenschaftlicher Forschung nicht adäquat.

Hinsichtlich der *Verkürzung* bedeutet dies, dass Validität ein wesentliches Kriterium für Modelle sein und auf allen Ebenen ihrer Konstruktion als Kriterium berücksichtigt werden muss. Daraus lässt sich eine Favorisierung von sog. *White-box*-Algorithmen ableiten, bei denen die Konstruktionsleistung des

30 Vgl. hierzu ausführlicher Scharloth (2018).

31 Vgl. Stachowiak (1973) 131–134.

32 Vgl. Geertz (1972) 9.

Algorithmus und die Auswirkungen der Veränderung von Parametern transparent und nachvollziehbar sind.[33]

Im Hinblick auf den *Pragmatismus* von Modellen muss im Gedächtnis behalten werden, dass ein einziges Modell für alle Anwendungsfälle nicht das Ziel des Forschungs- und Entwicklungsprozesses sein kann. Vielmehr variieren Algorithmen und Konstruktbestimmung mit der jeweiligen Forschungsfrage.

N-Gramm-basierte Paraphrasensuche

Ausgehend von diesen methodologischen Überlegungen haben wir uns für eine Modellierung des Konstrukts Paraphrase entschieden, die sich einerseits flexibel für unterschiedliche Forschungsfragen anpassen lässt (Pragmatizität), die andererseits aber hinreichend transparent und deren Validität damit auch überprüfbar ist: Wir modellieren eine Paraphrase als Ähnlichkeit komplexer n-Gramme.

N-Gramme sind Einheiten, die aus n Elementen bestehen. Normalerweise werden n-Gramme als Folge von Wortformen verstanden. Im Rahmen einer n-Gramm-Analyse werden alle im Korpus vorkommenden n-Gramme berechnet, wobei bestimmte Parameter wie Länge der Mehrworteinheit (aus zwei, drei oder mehr Wörtern bestehend) oder Spannweite (sind Lücken zwischen den Wörtern erlaubt?) festgelegt werden.[34]

Eine nur Oberflächenstrukturen berücksichtigende n-Gramm-Analyse ist jedoch kaum geeignet, Paraphrasen in all ihren oben skizzierten Facetten abzudecken. Daher betrachten wir nicht nur Wortformen als potentielle Einheiten von n-Grammen, sondern weitere interpretative linguistische Kategorien. Dies können zum einen Elemente sein, die sich auf die Tokenebene beziehen und die Wortform funktional oder semantisch deuten (als Repräsentant einer Wortart oder als Teil einer semantischen Klasse); zum anderen aber auch Elemente, die über die Tokenebene hinausgreifen, etwa das Tempus oder die Modalität einer Äußerung (direkte vs. indirekte Rede).

Welche Elemente welcher interpretativer Dimensionen in die Analyse mit einbezogen werden, hängt einerseits von der jeweiligen Forschungsfrage ab, andererseits forschungspraktisch auch davon, welche Ressourcen für die Annotation des Korpus zur Verfügung stehen. Das folgende Beispiel soll die Vorgehensweise bei einer komplexen n-Gramm-Analyse und die Effekte der Festlegung unterschiedlicher Parameter illustrieren.

33 Vgl. Rieder / Röhle (2012).

34 Vgl. Bubenhofer (2009) 149 ff.

Tabelle 1. Beispielsatz mit paradigmatischen und syntagmatischen Annotationen

	Angela	**Merkel**	**ist**	**die**	**erfolg-reichste**	**Regierungs-chefin**	**in**	**Europa**	**.**
Lemma	Angela	Merkel	sein	d	erfolg-reich	Regierungs-chefin	in	Europa	.
POS	NE	NE	VAFIN	ART	ADJA	NN	APPR	NE	$.
NP	[Angela	Merkel]		[die	erfolg-reichste	*Regierungs-chefin]	[in	*Europa]	
NER	[NE	Person]						NE Ort	
Hyperonym					gut, positiv	Regierungs-beamter, Verantwortli-cher, Leiter		Erdteil, Konti-nent	
Synonym					gelungen, sieghaft	Kanzler, Präsident		Abend-land, Okzident	

In unserem Beispiel stehen für den Prätext („Angela Merkel ist die erfolgreichste Regierungschefin in Europa.“) sowie für das (fiktive) Korpus von Paraphrasenkandidaten folgende interpretative Informationen zur Verfügung: Auf der Token-Ebene, also auf der Ebene von Wörtern und Satzzeichen, sind dies Informationen zum Lemma und zur Wortart des jeweiligen Token (Tokenisierung, Lemmatisierung und POS-Annotation erfolgten mittels des Tree-Taggers[35]), sowie unterschiedliche semantische Informationen, wie Hyperonyme (aus der semantischen Taxonomie GermaNet[36]) und Synonyme (aus Open Thesaurus[37]). Tokenübergreifend stehen Informationen zu sog. *Named Entities* und ihrer Klassifikation als Ort, Organisation, Person und anderes zur Verfügung, ebenso wie eine grundlegende Gliederung des Satzes in Teil-Nominalphrasen. Tabelle 1 gibt einen (um der Übersichtlichkeit willen vereinfachten) Überblick über die für die jeweiligen Token zur Verfügung stehenden Informationen.

Komplexe n-Gramme lassen sich nun einerseits durch syntagmatische Parameter wie die Festlegung von Grenzen (Länge des n-Gramms) und die Möglichkeit von Auslassungen, andererseits durch paradigmatische Ersetzungen erzeugen. Ein durch die Operationen der Begrenzung, Auslassung und Ersetzung erzeugtes

35 Vgl. Schmid (1995).
36 Vgl. Hamp / Feldweg (1997).
37 Vgl. Free Software Foundation (o.J.).

Tabelle 2. Beispielsatz mit paradigmatischen und syntagmatischen Annotationen und markierten paradigmatischen Ersetzungen (grau hinterlegte Felder) und Auslassungen (in grauer Schrift)

	Angela	**Merkel**	**ist**	**die**	**erfolg-reichste**	**Regierungs-chefin**	**in**	**Europa**	**.**
Lemma	Angela	Merkel	sein	d	erfolg-reich	Regierungs-chefin	in	Europa	.
POS	NE	NE	VAFIN	ART	ADJA	NN	APPR	NE	$.
NP	[Angela	Merkel]		[die	erfolg-reichste	*Regierungs-chefin]	[in	*Europa]	
NER	[NE	Person]						NE Ort	
Hyperonym					gut, positiv	Regierungs-beamter, Verantwortli-cher, Leiter		Erdteil, Konti-nent	
Synonym					gelungen, sieghaft	Kanzler, Präsident		Abend-land, Okzident	

n-Gramm ist eine abstrakte Repräsentation des Prätextes. Wie die n-Gramme jeweils gebildet werden sollen, wird durch die Festlegung der Parameter des Berechnungsalgorithmus gesteuert.

Wenn bei der Paraphrasensuche beispielsweise Personennamen durch die *named-entity*-Kategorie [NE Person] ersetzt werden, Verben als Wortform-Token erhalten bleiben, sämtliche Artikel und Adjektive ignoriert werden, alle Präpositionen durch ihren Part-of-Speech-Tag ersetzt werden und Substantive durch ihr Hyperonym (s. Tabelle 2), dann erhält man folgendes n-Gramm als Repräsentation des Prätextes: *[NE Person] ist [Regierungsbeamter, Verantwortlicher, Leiter] [APPR] [Erdteil, Kontinent].*

Die Paraphrasensuche besteht nun darin, in potentiellen Folgetexten nach Textstellen zu suchen, deren abstrakte Repräsentation mit der des Prätextes identisch ist, also nach Textstellen, die sich durch Begrenzung, Auslassung und Ersetzung ebenfalls auf das n-Gramm *[NE Person] ist [Regierungsbeamter, Verantwortlicher, Leiter] [APPR] [Erdteil, Kontinent]* zurückführen lassen. Dies wären beispielsweise Sätze wie „Putin […] ist […] Präsident […] von Russland." oder „Alfons der Viertelvorzwölfte […] ist […] König […] auf Lummerland." Das Beispiel zeigt, dass die gewählten Parameter den Algorithmus womöglich zu „gierig" machen, d. h. dass zahlreiche Textstellen gefunden werden, die nur schwerlich als Paraphrasen gelten können. Umgekehrt werden Sätze wie „Die erfolgreichste Regierungschefin in

Tabelle 3. Beispielsatz mit paradigmatischen und syntagmatischen Annotationen und markierten paradigmatischen Ersetzungen (grau hinterlegte Felder) und Auslassungen (in grauer Schrift)

	Angela	**Merkel**	**ist**	**die**	**erfolg-reichste**	**Regierungs-chefin**	**in**	**Europa**	**.**
Lemma	Angela	Merkel	sein	d	erfolg-reich	Regierungs-chefin	in	Europa	.
POS	NE	NE	VAFIN	ART	ADJA	NN	APPR	NE	$.
NP	[Angela	Merkel]		[die	erfolg-reichste	*Regierungs-chefin]	[in	*Europa]	
NER	[NE	Person]						NE Ort	
Hyperonym					gut, positiv	Regierungs-beamter, Verantwortli-cher, Leiter		Erdteil, Konti-nent	
Synonym					gelungen, sieghaft	Kanzler, Präsident		Abend-land, Okzident	

Europa ist Angela Merkel.“ mittels eines wie im Beispiel parametrisierten Algorithmus nicht gefunden. Eine Auflösung der syntagmatischen Struktur könnte hier zwar Abhilfe schaffen, würde den Algorithmus aber noch „gieriger“ machen. Das Hauptproblem des so konfigurierten Suchalgorithmus ist, dass der Name „Angela Merkel“ nicht als wesentlicher Bestandteil der Aussage im komplexen n-Gramm repräsentiert ist.

Führt man – wie in Tabelle 3 illustriert – mit dem Wortart-Tag NE annotierte Lexeme sowie Verben auf ihre lemmatisierte Form zurück, nimmt dagegen für Substantive und Adjektive die Hyperonyme als Elemente des n-Gramms auf und ersetzt Präpositionen durch den ihnen zugeordneten Tag, erhält man das n-Gramm *[Angela] [Merkel] [sein] [gut, positiv] [Regierungsbeamter, Verantwortlicher, Leiter] [APPR] [Europa]*, das nach Auflösung der syntagmatischen Struktur beispielsweise durch alphabetische Sortierung die Form *[gut, positiv] [sein] [APPR] [Angela] [Europa] [Merkel] [Regierungsbeamter, Verantwortlicher, Leiter]* hat.

Mit diesem komplexen n-Gramm kann spezifischer nach Aussagen über Angela Merkels Rolle in Europa gesucht werden. In Texten, in denen Namen exemplarische Funktionen haben, wie dies bei einigen pädagogischen aber auch philosophischen Genres der Fall sein kann, wäre eine so konfigurierte Suche allerdings weniger sinnvoll. Die Auflösung der syntagmatischen Struktur führt dazu, dass nicht nur Sätze wie „Die Kanzlerin Angela Merkel ist gut für Europa!“, sondern

auch Sätze wie „Für Europas Regierungschefs war Angela Merkel exzellent." als potentielle Paraphrasen identifiziert werden.

Das Beispiel zeigt, dass eine Suche von Paraphrasen mittels komplexer n-Gramme den Anforderungen an Modelle in den Digital Humanities genügt: Der Suchalgorithmus ist hochgradig konfigurierbar und transparent. In jedem Fall ist anhand der abstrakten Repräsentation von Prätext und Paraphrasenkandidat überprüfbar, aufgrund welcher sprachlichen Transformationen die beiden Textstellen als hinreichend ähnlich aufgefasst werden können, um als Paraphrase gedeutet zu werden. Entsprechend lässt sich auch der für die jeweilige Forschungsfrage zur Anwendung kommende operationale Begriff der Paraphrase anhand der interpretativen Auseinandersetzung mit den Ergebnissen der Suche datengeleitet verfeinern, ein Prozess, der auch das Potenzial hat, auf die Konstruktebene durchzuschlagen und den Begriff der Paraphrase im jeweiligen Forschungsfeld zu schärfen.

Doch selbst wenn eine für die jeweilige Forschungsfrage passende Konfiguration des Suchalgorithmus gefunden ist, werden vom Algorithmus zahlreiche Textstellen als Paraphrasenkandidaten identifiziert, die bei einer qualitativen Bewertung als nicht paraphrasenwertig klassifiziert werden können.

N-Gramm-basierte Paraphrasensuche in altgriechischen Texten

Will man die skizzierte Methode auf altgriechische Texte anwenden, so sieht man sich zunächst dem Problem gegenüber, dass computerlinguistische Ressourcen zur Anreicherung der Textkorpora mit interpretativen linguistischen Kategorien nicht in gleichem Umfang und gleicher Qualität zur Verfügung stehen wie für die großen lebenden Sprachen.

Insbesondere für die für die Paraphrasensuche zentrale Option der paradigmatischen Ersetzung von Token durch Synoyme oder Hyperonyme fehlt eine semantische Taxonomie. Im Rahmen des von der VW-Stiftung geförderten Projekts *Platon Digital* haben wir eine solche Taxonomie behelfsweise mit dem Arbeitstitel *Helleninet* erstellt. Dabei haben wir die semantischen Relationen zwischen Lexemen des Deutschen, wie sie im GermaNet[38] verzeichnet sind, auf das Altgriechische mittels maschineller Übersetzung der enthaltenen Lexeme übertragen.

Hierfür haben wir den Großteil der für uns erreichbaren digitalen Wörterbücher und Vokabellisten maschinell erschlossen und in ein verarbeitbares Format

38 Vgl. Hamp / Feldweg (1997).

gebracht. Einen Überblick in die verwendeten Ressourcen gibt Tabelle 4. Nur dann, wenn sich eine Lemma- und Übersetzungsrelation in mindestens zwei dieser Quellen fand, wurde sie in das *Helleninet* aufgenommen. Darüber hinaus haben wir eine Synonymenliste im Projekt erstellt und in die Taxonomie eingearbeitet.

Tabelle 4. Übersicht über die Ressourcen, die bei der Erstellung des *Helleninet* zum Einsatz kamen

Name der Ressource	**Anzahl Lexeme Altgriechisch**	**Anzahl Lexeme Deutsch**	**Bemerkung**
Operone	41925	55496	Onlinevokabelliste
Pape	21791	15489	Handwörterbuch der griechischen Sprache (1880)
Köbler	5467	6310	Abkunfts- und Wirkungswörterbuch (2007)
Synonymliste Leipzig	4278	764	Synonymliste der Arbeitsgruppe aus Leipzig
Altgriechisch.net	1831	2520	Onlinevokabelliste
Albertmartin	8076	10734	Onlinewörterbuch (Abfrage aller bekannten altgriechischen und deutschen Wörter / GermaNet-Einträge)
Glosbe	3910	3085	Onlinewörterbuch (Abfrage aller bekannten altgriechischen und deutschen Wörter / GermaNet-Einträge)
Gottwein	8388	15842	Onlinewörterbuch (Abfrage aller bekannten altgriechischen und deutschen Wörter / GermaNet-Einträge)

Diese Vorgehensweise ist aus sprachwissenschaftlicher Perspektive natürlich in hohem Maße problematisch. Einerseits haben sich in den Jahrtausenden sprachlicher Entwicklung Wortfelder und semantische Relationen vielfältig verschoben. Zum anderen führen polyseme Ausdrücke in einer Sprache zu Pseudorelationen in der anderen. Das Lemma βάκτρευμα etwa, das ins Deutsche als „Stab" übersetzbar ist, ist in GermaNet als polysemes Lexem einerseits als Artefakt, andererseits als Gruppe klassifiziert. Die Rückübersetzung der Kohyponyme aus GermaNet (also jener Wörter, die den gleichen Oberbegriff wie „Stab" haben) führt

dann dazu, dass neben den aus der Artefakt-Kategorie stammenden Lexemen (Fessel:δεσμός, Fessel:πέδη, Fessel:ἀρθροπέδη, Flügel:πτέρυξ, Gurt:ζώνη, Hebel:μοχλός, Leitung:ἀγωγή, Rohr:δονακογλύφος, Rohr:καλαμοφόρος, Rohr:κάλαμος, Schanze:ἀντιτείχισμα, Sitz:καθέδρα, Stock:βακτηρία, Verbindung:ἀρθμός, Verbindung:ἁρμονία, Walze:κύλινδρος, Wand:τοῖχος, Zelt:σκηνή, Zelt:σκήνωμα) auch jene aus der Gruppen-Kategorie in das *Helleninet* Eingang finden (Gemeinde:δῆμος, Generation:γενεά, Herde:βόσκημα, Herde:ἀγέλη, Kreis:γῦρος, Kreis:κύκλος, Kreis:τροχός, Menschheit:ἀνθρωπότης, Volk:δῆμος, Volk:λαός, Volk:λᾱοφθόρος, Volk:ἔθνος, Volksstamm:φυλή, Welt:γῆ).

Die geschilderte Vorgehensweise ist entsprechend nur als forschungspraktisch motivierte Notlösung zu betrachten. Gleichwohl enthält das maschinell erstellte *Helleninet* in der Mehrheit passende Synonymie- und Hyperonym-/Hyponymiebeziehungen, wie im Beispiel des Wortes καπνός, das als „Rauch" übersetzt werden kann. Das dazu in GermaNet verzeichnete Hyponym ist „Dampf", dessen altgriechische Übersetzungen (ἀτμίς, ἀτμός) auch als Hyponyme zu καπνός gelten können. Insgesamt enthält das *Helleninet* 6383 unterschiedliche altgriechische Lexeme, denen mindestens eine der 14 semantischen Relationen des GermaNet zugeordnet wird.

Bei der komplexen n-Gramm-Analyse in altgriechischen Texten wurden neben dem *Helleninet* der regelbasierte morphologische Parser und Lemmatisierer Morpheus[39] und der Mate-Tagger[40] kombiniert, um eine möglichst hohe Präzision bei der Bestimmung von Lemmata und Wortarten zu erreichen.[41]

Die folgenden exemplarischen Analysen wurden auf dem ‚Goldstandard'[42] durchgeführt. Dabei wurden sämtliche Prätexte mit sämtlichen Paraphrasen verglichen und zwar mittels zweier unterschiedlicher Konfigurationen des Suchalgorithmus.

In der ersten Analyse wurden sämtliche Tetragramme (also Vier-Wort-Einheiten) in einem Fenster von zwölf Token berechnet, wobei das Tetragramm mindestens zwei Inhaltswörter (Substantive, Adjektive, Verben) enthalten musste, in der zweiten sämtliche Pentagramme (Fünf-Wort-Einheiten) in einem Fenster von 14 Token, wobei jedes Pentagramm mindestens drei Inhaltswörter enthalten musste. Die syntagmatische Struktur wurde zugunsten einer alphabetischen Sortierung aufgelöst. Folgende paradigmatische Ersetzungen waren möglich: Artikel, Pronomen, Konjunktionen, Interjektionen, Adverbien, Partikel, Präpositionen und

39 Crane (1991).

40 Björkelund/Bohnet/Hafdell/Nugues (2010); online unter: <https://code.google.com/p/mate-tools>.

41 Zu den Problemen beim Tagging altgriechischer Texte vgl. Celano/Crane/Majidi (2016).

42 Vgl. den Beitrag in Appendix 1 „Ein Parallelkorpus von Paraphrasen auf Platon: Der ‚Goldstandard' des Projekts *Platon Digital*" in diesem Band S. 275.

Numerale konnten im n-Gramm durch den Wortarten-Tag ersetzt werden; Substantive, Adjektive, Verben und Pronomen konnten durch die lemmatisierte Form ersetzt werden; Substantive, Verben und Adjektive konnten auch durch *Helleninet*-Hyperonyme ersetzt werden.

Aus der Textstelle Ὁ δέ ἐστιν ὁ θάνατος χωρὶς εἶναι τὴν ψυχὴν τοῦ σώματος (Der Tod aber ist die Getrenntheit der Seele vom Leibe) aus Plotins *Enneaden* (I 6,6) lässt sich unter anderem das Wortformen-Pentagramm Ὁ δέ ἐστιν θάνατος ψυχὴν extrahieren[43], in dem sich beispielsweise Ὁ durch den Pronomen-Tag *:p:* und δέ durch den Partikel-Tag *:g:* ersetzen lassen; zum Substantiv ψυχή findet sich im *Helleninet* das Hyperonym ἄνθρωπος, so dass das nach UTF-8 Codepunkten sortierte Pentagramm *:d: :p: _ἄνθρωπος_ ψυχὴν ἐστιν* lautet. Durchsucht man mit diesem n-Gramm mit der gleichen Konfiguration die Platon-Texte des ‚Goldstandards', findet man die folgende Stelle aus dem *Phaidon*, die der Prätext der Plotin-Paraphrase ist:

> ἆρα μὴ ἄλλο τι ἢ τὴν τῆς ψυχῆς ἀπὸ τοῦ σώματος ἀπαλλαγήν; καὶ εἶναι τοῦτο τὸ τεθνάναι, χωρὶς μὲν ἀπὸ τῆς ψυχῆς ἀπαλλαγὲν αὐτὸ καθ' αὑτὸ τὸ σῶμα γεγονέναι, χωρὶς δὲ τὴν ψυχὴν [ἀπὸ] τοῦ σώματος ἀπαλλαγεῖσαν αὐτὴν καθ' αὑτὴν εἶναι; ἆρα μὴ ἄλλο τι ᾖ ὁ θάνατος ἢ τοῦτο; (Plat. *Phaid.* 64 c).[44]

Die unterschiedlichen Konfigurationen des Algorithmus führten in der Gesamtschau freilich zu unterschiedlichen Ergebnissen, wie Tabelle 5 zeigt. Als Gütekriterien für die Qualität der jeweiligen Konfiguration wurden zwei Werte berechnet: Der *Precision-Wert* ist der Quotient aus der Anzahl der gefundenen Paraphrasen und der Anzahl aller als Paraphrasen identifizierten Dokumente (also einschließlich der fälschlicherweise als Paraphrasen klassifizierten Dokumente). Die Präzision der Paraphrasensuche wird also über den Anteil der richtig als Paraphrasen klassifizierten Texte an allen als Paraphrasen klassifizierten Texten bestimmt. Der *Recall-Wert* ist der Quotient der Anzahl der gefundenen Paraphrasen und der Anzahl aller Paraphrasen im Korpus. Er sagt aus, welchen Anteil der tatsächlichen Paraphrasen mit Hilfe des jeweiligen Algorithmus gefunden wurden, und lässt

43 N-Gramme bilden nicht notwendigerweise den Satzinhalt ab – in den allermeisten Fällen enthalten sie lediglich sprachliche Fragmente. Bei der komplexen n-Gramm-Analyse werden systematisch alle möglichen Kombinationen von n Elementen im Satz berechnet. Die für das Beispiel gewählte Kombination ist eine zufällige Auswahl aus der großen Anzahl von Kombinationen.

44 ‚Und wohl etwas andres als die Trennung der Seele von dem Leibe? Und daß das heiße tot sein, wenn abgesondert von der Seele der Leib für sich allein ist und auch die Seele abgesondert von dem Leibe für sich allein ist? Oder sollte wohl der Tod etwas anderes sein als dieses?' (Übers. Schleiermacher); vgl. Appendix 1, Nr. 79.

dabei die fälschlicherweise als Paraphrasen klassifizierten Dokumente außer Acht. Je näher beide Werte zum Wert 1 liegen, desto besser.

Tabelle 5. *Precision*- und *Recall*-Werte der Paraphrasensuche mit unterschiedlichen Parametern

	Precision		**Recall**	
False (but resonable) positive	**Einbezogen**	**Nicht einbezogen**	**Einbezogen**	**Nicht einbezogen**
Fenster 12, 2 Inhaltswörter, Tetragramme	0,676	0,497	0,788	0,821
Fenster 14, 3 Inhaltswörter, Pentagramme	0,821	0,707	0,381	0,248

Bei der qualitativen Analyse der Ergebnisse wurden die sog. *false positives,* also Dokumente, die zwar als Paraphrasen identifiziert wurden, jedoch keine waren, in zwei Klassen eingeteilt: solche, die keine plausible Interpretation als Paraphrase zuließen, und solche, bei denen die Kategorisierung als Paraphrase zwar plausibel, die jedoch aus Kenntnis des Co-Textes und der Rezeptionsgeschichte nicht als Paraphrasen infrage kamen. *Precision* und *Recall* wurden auf der Basis dieser Klassifikation zweimal berechnet: Einmal wurden die plausiblen Paraphrasen in die Liste der Treffer mit einbezogen, einmal wurden sie als fehlerhafte Klassifikationen gewertet. Dahinter stand das Kalkül, dass in künftigen Paraphrasensuchen plausible Treffer nicht ausgeschlossen werden sollten.

Im Ergebnis liegen zwar die *Precision*-Werte bei Pentagrammen und einem höheren Anteil von Inhaltswörtern höher als bei Tetragrammen, was bedeutet, dass es sich bei den gefundenen Textstellen häufiger tatsächlich um Paraphrasen handelt. Der *Recall*-Wert ist allerdings mit 0.381 vergleichsweise schlecht, bedeutet er doch, dass weniger als die Hälfte aller Paraphrasen überhaupt gefunden wurden. Bei der Paraphrasensuche mittels Tetragrammen liegt der *Recall*-Wert immerhin bei 0.788, allerdings ist der *Precision*-Wert lediglich bei 0.676. Die höhere Trefferquote wird also mit einer größeren Anzahl an *false positives* erkauft.

N-Gramm-basierte Paraphrasensuche am Beispiel der Presselenkung im Nationalsozialismus

Die Paraphrasensuche kann aber nicht nur dazu eingesetzt werden, intentionale und damit freiwillige Übernahmen von kanonischen Prätexten zu identifizieren, sondern auch den Grad des Einflusses von angeordneten inhaltlichen Übernahmen aus Vorgaben, wie dies bei Presseanweisungen in Diktaturen der Fall ist. Die Paraphrasenanalyse kann hier dazu genutzt werden, Art und Grad der Durchherrschung zu messen. Dies soll im Folgenden am Beispiel der Aufnahme nationalsozialistischer Tagesparolen im *Völkischen Beobachter* (Wiener Ausgabe) geschehen.

Der Terminus der Durchherrschung wurde von Jürgen Kocka mit Blick auf das gesellschaftliche und politische System der DDR geprägt und bezeichnet eine „ubiquitäre poltische Herrschaft", die eine Gesellschaft „bis in ihre feinsten Verästelungen" prägt.[45] Hachtmann übertrug den Terminus auf das Dritte Reich, in dem er eine kumulative Durchherrschung ausmacht, weil sie „polykratisch unabgesprochen" geschah.[46] Ein wichtiger Baustein dieser Herrschaft war freilich die zentralisierte Lenkung von Informationen. Um die Informationen, die die Bevölkerung erhalten sollte, inhaltlich zu vereinheitlichen und zu steuern, aber auch zur Durchsetzung bestimmter Regelungen der propagandistischen Sprache gab das Reichsministerium für Volksaufklärung und Propaganda sog. „Anweisungen der Pressekonferenz der Reichsregierung des Dritten Reichs" heraus. Sie wurden auf einer täglich stattfindenden Pressekonferenz verkündet und auch als Protokolle den Redaktionen der verbliebenen Zeitungen zur Verfügung gestellt.[47]

Als die Anzahl der Presseanweisungen so groß wurde, dass ihre Befolgung und Hierarchisierung für Redaktionen zunehmend schwierig wurden, richtete das Reichsministerium für Volksaufklärung und Propaganda die Tagesparolen als weitere Ebene für die Steuerung der Presse ein. Der Leiter der Abteilung *Deutsche Presse* Hans Fritzsche erklärte bei ihrer Einführung am 31. 10. 1940, die Tagesparole enthalte „alles, was fuer die Presse verbindlich sei"; in ihr würden die „taeglichen politischen Weisungen auf die kuerzeste Formel gebracht".[48] Täglich gab das Propagandaministerium drei bis fünf Tagesparolen heraus.[49] Anders als für die Vorkriegspresseanweisungen[50] liegt für die Tagesparolen leider keine vollständige

45 Kocka (1994) 548.
46 Hachtmann (2011) 478.
47 Vgl. Schmitz-Berning (2010).
48 Zitiert nach Wilke (2007) 228.
49 Vgl. Wilke (2007) 229.
50 Vgl. Bohrmann / Toepser-Ziegert (1984–2001).

Edition vor. Für die Analyse musste daher auf die äußerst selektive Edition von Sündermann[51] zurückgegriffen werden.

Um die Wirkung der NS-Presselenkung zu messen, sollen die Ausgaben des *Völkischen Beobachters* daraufhin untersucht werden, ob sich in ihnen Paraphrasen der Tagesparolen finden. Auch hier stellte sich allerdings das Problem, dass der *Völkische Beobachter* wie andere NS-Zeitungen zwar an unterschiedlichen Orten als Digitalisat vorgehalten wird, jedoch wegen des Verbots der Verbreitung volksverhetzender Inhalte nicht über das Internet zugänglich ist; auch die Zugänge an lokalen Standorten erlauben nur begrenzte Exporte von Texten. Einzig die Österreichische Nationalbibliothek macht eine größere Menge der Wiener Ausgabe des *Völkischen Beobachters* frei online verfügbar.[52] Doch auch hier gibt es forschungspraktische Hürden: Zwar lassen sich die mittels OCR in Text umgewandelten Digitalisate maschinell aus dem Internet laden, allerdings ist die Qualität der OCR für die in Fraktur gedruckten Ausgaben so miserabel, dass sie für maschinelle Sprachanalysen nicht geeignet ist. Von den auf den Seiten der Österreichischen Nationalbibliothek zur Verfügung stehenden Ausgaben sind erst jene ab März 1944 in Antiqua gesetzt. Hier ist die Qualität der OCR auf einem brauchbaren Niveau.

Die folgenden Analysen wurden daher unter einer doppelten Einschränkung durchgeführt: Sie umfassen einen Zeitraum von fünf Monaten von März bis einschließlich Juli 1944 und sie beschränken sich notwendig auf jene 46 Tagesparolen, die sich für diesen Zeitraum in der Edition von Sündermann finden.

Die Korpora wurden mit dem *TreeTagger*[53] lemmatisiert und mit Part-of-Speech-Informationen annotiert. Zusätzlich wurden Komposita mit Hilfe von *Morphisto*, das auf dem SFST-Toolkit beruht, und mit der morphologischen Komponente SMOR[54] in ihre lexikalischen Bestandteile zerlegt. Die 126 Ausgaben des *Völkischen Beobachters* enthalten 4.299.072 laufende Wortformen, die 46 Tagesparolen 3.500.

Die Analyse erfolgte mittels zweier unterschiedlicher Konfigurationen des Suchalgorithmus: In einem Fall wurden Trigramme aus Inhaltswörtern in einem Fenster von acht Token gebildet, im anderen Bigramme aus Inhaltswörtern in einem Fenster von zwölf Token, wobei in beiden Konfigurationen die Inhaltswörter durch Hyperonyme oder Lemmata ersetzt werden konnten. Die maschinell ermittelten Paraphrasenkandidaten aus beiden Analysen wurden interpretativ evaluiert. Tabelle 6 gibt einen vergleichenden Überblick über die Fundstellen für Paraphrasen.

---

51 Vgl. Sündermann (1973).

52 Vgl. <http://anno.onb.ac.at/info/vob_info.htm> und <http://anno.onb.ac.at/cgi-content/anno?aid=vob>.

53 Vgl. Schmid (1994; 1995).

54 Vgl. Schmid / Fitschen / Heid (2004).

Tabelle 6. Übersicht über die Fundstellen der Tagesparolen-Paraphrasen im *Völkischen Beobachter*

Tagesparole	Bigramm-Suche	Trigramm-Suche
01.03.44	02.03.1944, S. 2 05.03.1944, S. 2	02.03.1944, S. 2 05.03.1944, S. 2
20.03.44	20.03.1944, S. 1 21.03.1944, S. 2 22.03.1944, S. 2	20.03.1944, S. 1 21.03.1944, S. 2 22.03.1944, S. 2
06.04.44	07.04.1944, S. 1 09.04.1944, S. 1	07.04.1944, S. 1 09.04.1944, S. 1
14.04.44	16.04.1944, S. 2 18.04.1944, S. 2	16.04.1944, S. 2 18.04.1944, S. 2
24.04.44		
04.05.44	04.05.1944, S. 1	04.05.1944, S. 1
09.05.44	12.05.1944, S. 1	12.05.1944, S. 1
11.05.44	13.05.1944, S. 5 15.05.1944, S. 1, 2	13.05.1944, S. 5 15.05.1944, S. 1
13.05.44	14.05.1944, S. 1 15.05.1944, S. 1	15.05.1944, S. 1
16.05.44	17.05.44, S. 1	17.05.44, S. 1
18.05.44		
20.05.44		
23.05.44		
27.05.44		
01.06.44	04.06.1944, S. 2	
05.06.44	06.06.1944, S. 1	06.06.1944, S. 1
06.06.44	07.06.1944, S. 1, 2 08.06.1944, S. 1, 5 10.06.1944, S. 2	07.06.1944, S. 1, 2 08.06.1944, S. 5
07.06.44	11.06.1944, S. 1	11.06.1944, S. 1
08.06.44	09.06.1944, S. 1 10.06.1944, S. 1	09.06.1944, S. 1 10.06.1944, S. 1
10.06.44		
10.06.44	11.06.1944, S. 1	
13.06.44	17.06.1944, S. 1	17.06.1944, S. 1
16.06.44	20.06.1944, S. 2	
16.06.44	17.06.1944, S. 1, 2 19.06.1944, S. 2	17.06.1944, S. 1, 2 19.06.1944, S. 2

Tabelle 6. *(Fortsetzung)*

Tagesparole	Bigramm-Suche	Trigramm-Suche
17.06.44	18.06.1944, S. 1, 2 19.06.1944, S. 1 20.06.1944, S. 2	18.06.1944, S. 1, 2 19.06.1944, S. 1 20.06.1944, S. 2
19.06.44	20.06.1944, S. 1 21.06.1944, S. 1	21.06.1944, S. 1
19.06.44		
19.06.44	21.06.1944, S. 1	21.06.1944, S. 1
20.06.44	21.06.1944, S. 1 23.06.1944, S. 1 24.06.1944, S. 1	21.06.1944, S. 1 23.06.1944, S. 1 24.06.1944, S. 1
26.06.44	27.06.1944, S. 1	27.06.1944, S. 1
28.06.44	29.06.1944, S. 1, 2 02.07.1944, S. 1	29.06.1944, S. 1, 2 02.07.1944, S. 1
01.07.44	02.07.1944, S. 1 03.07.1944, S. 1	02.07.1944, S. 1 03.07.1944, S. 1
01.07.44	02.07.1944, S. 2 04.07.1944, S. 2	02.07.1944, S. 2
02.07.44	03.07.1944, S. 1	03.07.1944, S. 1
03.07.44	05.07.1944, S. 2	
03.07.44	04.07.1944, S. 3	04.07.1944, S. 3
04.07.44	05.07.1944, S. 1, S. 2 06.07.1944, S. 2 07.07.1944, S. 1 08.07.1944, S. 2	05.07.1944, S. 1 07.07.1944, S. 1
06.07.44	07.07.1944, S. 1	
08.07.44		
08.07.44		
12.07.44		
15.07.44		
19.07.44		
21.07.44	22.07.1944, S. 1 23.07.1944, S. 1 24.07.1944, S. 1	22.07.1944, S. 1 23.07.1944, S. 1 24.07.1944, S. 1
26.07.44	27.07.1944, S. 1, 2 28.07.1944, S. 1, 2 29.07.1944, S. 1, 2 30.07.1944, S. 2	27.07.1944, S. 1, 2 28.07.1944, S. 1, 2 29.07.1944, S. 1, 2 30.07.1944, S. 2
31.07.44		

Im Hinblick auf die Frage, ob und in welcher Weise die Tagesparolen einen Einfluss auf die Berichterstattung des *Völkischen Beobachters* hatten, kann im Einklang mit früherer Forschung[55] festgestellt werden, dass der Grad der inhaltlichen Lenkung erheblich war. Dies zeigt sich daran, dass ein Großteil der Tagesparolen im *Völkischen Beobachter* zumindest ausschnittweise paraphrasiert wiedergegeben wurde. Jene Tagesparolen, für die sich keine Paraphrasen finden ließen, enthalten größtenteils Berichterstattungsverbote[56] oder Hinweise auf globale Tendenzen im Kriegsgeschehen und den Verweis auf Berichte des Oberkommandos der Wehrmacht[57]. Darüber hinaus finden sich, von wenigen Ausnahmen abgesehen, fast alle Paraphrasen auf den ersten beiden Seiten. Schließlich ist auch bemerkenswert, dass einzelne Tagesparolen sich nicht nur auf die Berichterstattung an einem Tag, sondern auch an Folgetagen auswirkten.

Anhand der Intensität der Paraphrasierung lässt sich zudem abschätzen, welche Tagesparolen als besonders wichtig erachtet wurden. Dazu zählt beispielsweise die Tagesparole vom 4. 7. 1944, in der der Einsatz der V 1 als Vergeltung für den Bombenterror gerechtfertigt und auf die Reaktionen der britischen Öffentlichkeit verwiesen wird.[58] Von noch größerer Bedeutung war die Tagesparole zum Erlass Adolf Hitlers über den verstärkten totalen Kriegseinsatz vom 26. 7. 1944,[59] der die Berichterstattung des *Völkischen Beobachters* über mehrere Tage hin dominierte.

Vergleicht man die beiden Konfigurationen des Suchalgorithmus im Hinblick auf die Anzahl der Fundstellen, so wird deutlich, dass die Suche mittels Trigrammen kaum weniger Treffer liefert als die Suche mittels Bigrammen. Dennoch erweist sich der „gierigere" Suchalgorithmus insbesondere bei weniger prominent verhandelten Themen als treffsicherer, etwa im Fall der Tagesparole zur Rede des finnischen Ministerpräsidenten Edwin Linkomies vom 3. 7. 1944, die lediglich auf Seite 2 verhandelt wurde.[60]

---

55 Vgl. Wilke (2007) und Dussel (2010).

56 Beispielsweise die Tagesparole vom 8. 7. 1944: „Die Bayreuther Festspiele dürfen auch in diesem Jahr weder in der Reichspresse noch in der Lokalpresse vor ihrem Abschluß auf irgend eine Weise erwähnt werden. Dies gilt nicht nur für den redaktionellen Teil, sondern auch für die Aufnahme von Anzeigen jeglicher Art, in denen sich irgend ein – wenn auch indirekter Hinweis auf die Festspiele oder gar die Zeit ihrer Abhaltung – befindet." (Zitiert nach Sündermann [1973] 275.)

57 Etwa die Tagesparole vom 19. 7. 1944: „Die Abwehrkämpfe stehen im Osten und Westen weiterhin im Zeichen starker feindlicher Angriffe, denen die deutschen Truppen mit größter Entschlossenheit begegnen. Bei der Behandlung des OKW-Berichtes sind diese Gesichtspunkte hervorzuheben." (Zitiert nach Sündermann [1973] 276.)

58 Vgl. Sündermann (1973) 272.

59 Vgl. Sündermann (1973) 277.

60 Sie lautet: „Die Linkomies-Rede verdient in der deutschen Presse auch aufmachermäßig in besonderer Weise hervorgehoben und unterstrichen zu werden. Wie schon die Führer-Rede, so legt auch diese Kundgebung ein Zeugnis dafür ab, daß es für ein Volk, das in dem

Allerdings ergibt die Bigramm-Suche auch zahlreiche Übereinstimmungen mit Textstellen, die nicht ohne Weiteres als Paraphrasen zu deuten sind. Dies betrifft insbesondere Formulierungsvorlieben in der NS-Propagandasprache, die als konventionalisierte Muster auf Bigramme zurückzuführen sind wie „erbitterter Widerstand" / „harter Widerstand" / „harte Kämpfe" / „schwere Kämpfe" / „schwerste Kämpfe" oder „entschlossene Haltung" / „entschiedene Haltung" oder „stolze Haltung" / „vorbildlicher Haltung" / „tadellose Haltung" oder „größte Bewährungsprobe" / „harte Probe" / „härteste Probe".

Auch das Ideologem vom Überlebenskampf des von gezielter Ausrottung bedrohten deutschen Volkes (und anderer Völker) wird im Korpus an unterschiedlichsten Stellen gefunden, selbst dort, wo es nicht als Paraphrase der Tagesparole vom 3. 7. 1944 gedeutet werden kann, in der es heißt: „Die Lehre aus der Entwicklung in Finnland ist geeignet, als ein Beispiel dafür herangezogen zu werden, daß ohne Kampf bis zum Äußersten die nationale Freiheit und Existenz eines europäischen Volkes gegenüber dem bolschewistischen Vernichtungswillen nicht mehr gewahrt werden können. Dieser kalten und nackten Tatsache gegenüber entlarvt sich jede demokratische Phraseologie selbst als gefährliche Wegbereiterin zur Vernichtung der europäischen Völker."[61] Im *Völkischen Beobachter* ist im Untersuchungszeitraum teilweise mehrfach von „Ausrottung des weißruthenischen Volkes", „Ausrottung der weißruthenischen Bevölkerung", „Ausrottung unseres Volkes", „Ausrottung der nicht russisch-bolschewistischen Völker", „Ausrottung der zivilen Bevölkerung", „Ausrottung des deutschen Volkes", „Ausrottung ganzer Völker", „Ausrottung unseres Volkes", „Vernichtung des deutschen Volkes", „Vernichtung aller nichtbolschewistischen Völker", „Vernichtung des deutschen Volkes", „Vernichtung unseres Volkes", „Vernichtung der Völker", „Vernichtung der nordkaukasischen Stämme", „Vernichtung dieser Völker" die Rede.

So verweist die „gierigere" Konfiguration des Suchalgorithmus über den Einzeltextbezug hinaus auf die der nationalsozialistischen Propaganda zugrundeliegende semantische Tiefenstruktur, auf kontextabstrakte Topoi.

heutigen großen Entscheidungskampf seine Zukunft sichern will, keinen anderen Weg gibt, als unter allen Umständen um Leben und Ehre zu kämpfen." (Zitiert nach Sündermann [1973] 274.)

61 Zitiert nach Sündermann (1973) 274 f.

Fazit

Ziel dieses Beitrags war es, eine zum interpretativen und rekursiven Denkstil der Digital Humanities passende Vorgehensweise für die Suche von Paraphrasen vorzustellen. Dabei sind wir davon ausgegangen, dass die Forschung in den Digital Humanities stets von einer doppelten Unschärfe geprägt ist, nämlich sowohl hinsichtlich der untersuchten Konstrukte als auch hinsichtlich ihrer empirischen Erscheinungsformen. Weil konzeptuelle und empirische Erkenntnisweise nicht von einander zu trennen sind, favorisieren wir den Einsatz von Algorithmen, die gleichermaßen konfigurierbar und transparent sind. Für die Suche nach Paraphrasen, für die die wechselseitige Transformierbarkeit von einzelnen Textfragmenten zum definitorischen Kern gehört, halten wir daher komplexe n-Gramme für eine geeignete maschinelle Operationalisierung.

Mit Hilfe der Analysen konnten wir zeigen, dass Suchalgorithmen, die auf komplexen n-Grammen beruhen, auf so unterschiedliche Weise konfiguriert werden können, dass sie für die ganze Bandbreite der mit dem Forschungsgegenstand der Paraphrase in den Blick kommenden Forschungsfragen als Hilfsmittel dienen können. So konnten in den exemplarischen Analysen sowohl Spuren von Einzeltextbezug, als auch von Systembezug mit unterschiedlichen Konfigurationen des Suchalgorithmus aufgespürt werden.

Insbesondere bei der Suche nach Systembezügen, die „gierigere" Konfigurationen des Suchalgorithmus erfordern, kann freilich das Problem auftreten, dass sehr viele *false positives* gefunden werden. Zudem steht und fällt die Qualität der Paraphrasensuche mit der Qualität der zur Verfügung stehenden computerlinguistischen Werkzeuge zur Anreicherung der Korpusdaten mit interpretativen linguistischen Informationen. Für das Altgriechische sind diese Hilfsmittel und die von uns entwickelten Workarounds noch nicht von befriedigender Qualität. Unsere Analysen haben wie die anderen Beiträge in diesem Band dennoch gezeigt, dass eine Paraphrasensuche in altgriechischen Texten zwar dornen-, aber dennoch erfolgreich sein kann.