Ann-Marie Moser und Stefano De Pascale*

Klassische Textsortengliederung und quantitativ ermittelte Textgruppen in der frühen Neuzeit: ein Vergleich

Traditional text classifications and computationally derived text groupings in corpora of Early New High German: a comparison

https://doi.org/10.1515/zgl-2025-2019

Abstract: This study explores the relationship between manual text classifications linked to corpus metadata and computationally derived text groupings in corpora of Early New High German. We employ vector space models which can arrange texts in a multidimensional space based on semantic similarities and then cluster 463 texts from the ReF and GerManC corpora (1350–1800) into lexically and semantically defined groups. This operationalization of text types allows for the observation that there are more or less prototypical representatives of a text type and that there are overlaps and divergences in the development of such types. We evaluate the result of our quantitative analysis in a LASSO regression model which predicts the relative frequency of wh-relative pronouns, a linguistic variable known for genre-sensitive and historical variation. Our results show that data-driven clustering are at least complementary to traditional classifications in capturing semantic distinctions and diachronic variation in textual traditions. The findings contribute to historical text linguistics by proposing a bottom-up methodology for identifying text types and revealing how genre evolution correlates with linguistic change.

***Kontaktpersonen: Dr. Ann-Marie Moser**: Universität Zürich, Deutsches Seminar, Schönberggasse 9, CH-8001 Zürich, doppelte Affiliation: Linguistik Zentrum Zürich, Andreasstrasse 15, CH-8050 Zürich, E-Mail: ann-marie.moser@ds.uzh.ch. https://orcid.org/0000-0002-8003-0722
Prof. Dr. Stefano De Pascale: Vrije Universiteit Brussel, Brussels Centre for Language Studies (BCLS), Pleinlaan 2, BE-1050 Elsene, E-Mail: stefano.de.pascale@vub.be. KU Leuven, Quantitative Lexicology and Variational Linguistics (QLVL), Blijde-Inkomststraat 21 bus 3308, BE-3000 Leuven, E-Mail: stefano.depascale@kuleuven.be. https://orcid.org/0000-0003-2455-9004

1 Einleitung

Die frühe Neuzeit[1] ist bekannt als eine Epoche tiefgreifender technischer, gesellschaftlicher und sprachlicher Veränderungen. Der Durchbruch zu einer Schriftkultur auf Deutsch (vgl. Kuhn 1969: 264, Betten 1987: 20–22) geht einher mit einer Zunahme an Texten und der (Weiter)Entwicklung von Textsorten und Texttraditionen aus verschiedenen Bereichen wie dem Weltlichen, Alltäglichen, Fachlichen, Erbaulichen und Unterhaltenden sowie generell in der Prosa (vgl. Polenz 2000: 114, Betten 1987, 57–62).[2] Textsorten sind das Resultat gesellschaftlich verfestigter Muster und kommunikativer Prozesse (vgl. Günthner/Knoblauch 1994: 695–696). Komplett

1 Unter dem Begriff „frühe Neuzeit" wird der Zeitraum von 1350 bis 1800 verstanden: Er umfasst die Periode des Frühneuhochdeutschen (1350 bis 1650) und geht weiter bis ungefähr 1800. In diesen Zeitraum fallen u. a. die Entstehung des Bürgertums, weitreichende Veränderungen in der Medien- und Textsortenlandschaft (vgl. Elspaß 2008: 6; Polenz 2000: 103) sowie von der Mitte des 16. bis Ende des 18. Jahrhunderts eine vom Absolutismus geprägter Zeitraum (vgl. Polenz 2013: 1). In der Forschungsliteratur finden sich Schreibungen sowohl mit Majuskel (**F**rühe Neuzeit) (z. B. bei Elspaß 2008, Solms 2000) als auch mit Minuskel (**f**rühe Neuzeit) (z. B. bei Besch 2003, Polenz 2000: 103) und nicht immer ist klar definiert, welcher Zeitraum unter diesem Begriff verstanden wird. Wir haben uns hier für die Kleinschreibung entschieden, da die „frühe Neuzeit" – zumindest in der Sprachwissenschaft – (bisher) kein feststehender, klar umrissener Begriff ist, siehe z. B. das *Lexikon der Germanistischen Linguistik* (Althaus et al. 1980), das über keinen Eintrag unter „Neuzeit" bzw. „frühe Neuzeit" verfügt.

2 Bisher fehlt ein umfassender Überblick zur Textsortengeschichte des Deutschen und/oder der frühneuhochdeutschen Epoche. Für die frühe Neuzeit seit daher verwiesen auf die folgenden Werke: Die *Deutsche Sprachgeschichte* von Polenz, Band 1 (2000) und 2 (2013). In Band 1 v. a. Kap. 4.2, in Band 2 v. a. Kap. 5.2. Im HSK *Sprachgeschichte* gibt es zudem thematisch einschlägige Beiträge wie jenen zu Kanzleisprachen (Bentzinger 2000) oder zu den soziokulturellen Voraussetzungen des Frühneuhochdeutschen von Solms (2000). Darüber hinaus finden sich auch im Sammelband von Haaf/Schuster (2023) zahlreiche Beiträge zu Textmusterwandel in der frühen Neuzeit.

neue Textsorten entstehen im Zeitraum von 1350 bis 1800 selten, der Regelfall sind vielmehr Konvergenzen (Überlappungen) und Divergenzen (Ausdifferenzierungen) von Texttraditionen (vgl. Schuster 2019: 230). Oesterreicher (1997: 21) spricht sich dafür aus, den Begriff der „Diskurstradition" dem Begriff der Textsorte oder Texttyp vorzuziehen, da dieser „die Konventionalität, mithin die notwendige Historizität der genannten Muster und Schemata, schon in der Bezeichnung zum Ausdruck bringt" (Oesterreicher 1997: 21). Der Begriff „Diskurstraditionen" ist vor allem in der romanistischen Tradition bekannt (vgl. z. B. Kabatek 2015), in der germanistischen Tradition wird vielfach der Begriff der „Texttraditionen" verwendet (vgl. z. B. Schuster 2019, Schoenke 2000).

Neben der frühen Neuzeit befasst sich die historische Textlinguistik auch mit Texttraditionen im Alt- und Mittelhochdeutschen sowie im (wahrscheinlich) textsortenreichen 19. Jahrhundert.[3] Grundsätzlich ist zu beobachten, dass sich die Ausbildung von Textsorten in der Geschichte des Deutschen nicht mit den üblichen Epochengliederungen und den angenommenen Schritten hin zur Entstehung einer schriftlichen Standardsprache parallelisieren lässt (vgl. Schuster 2019: 229). Auch wenn sich nur selten Tradierungslinien vom Althochdeutschen bis zur heutigen Zeit nachweisen lassen (vgl. z. B. Sonderegger 1979: 32), sind dennoch allgemeine Tendenzen der Textsortenentwicklung ersichtlich: Ein Beispiel hierfür stellt die Loslösung von vorwiegend lateinischen Texttraditionen vom Althochdeutschen bis in die heutige Zeit dar (vgl. Schuster 2019: 229–230).

Textsortenklassifikationen beruhen auf einer sorgfältigen philologischen Analyse, häufig steht dabei die Autorin/der Autor mit ihren/seinen bestimmten zeitlichen, räumlichen, sozialen usw. Bindungen im Mittelpunkt (darauf beruht bspw. das frühneuhochdeutsche Lesebuch von Reichmann/Wegera 1988). Seitdem zunehmend größere Korpora digital zur Verfügung stehen, hat auch die Zahl korpusbasierter oder korpusgetriebener Arbeiten zur Textsortenbetrachtung zugenommen. Ein aktuelles Beispiel für diese Richtung ist der von Haaf/Schuster (2023) herausgegebene Band zu historischen Textmustern im Wandel sowie der Beitrag von Mazzola et al. (2023) zu Text- und Diskurstraditionen im Spanischen. Unser Beitrag knüpft an diese Forschungsrichtung an; er baut insbesondere an dem Beitrag von Mazzola et al. (2023) auf und entwickelt den dort verwendeten Ansatz zur quantitativen Modellierung von Texttraditionen weiter. Dieser Beitrag geht der Frage nach, ob „Textsorten" bottom-up induziert werden können anstatt sie mit Hilfe von Expertenwissen (im Sinne einer philologischen Analyse) zu bestimmen.

3 Das 19. Jahrhundert ist erst ansatzweise erforscht, vgl. Schuster (2019: 235), die es als „Forschungsdesiderat" bezeichnet. Siehe aber Thielert/Georgl (2023), Schuster et al. (2023) und Hausendorf (2023), die sich mit Textsorten der Pressekommunikation im 19. Jahrhundert befassen.

„Textsorten“ haben wir an dieser Stelle bewusst in Anführungszeichen gesetzt, denn mithilfe unserer Methode (für Details s. v. a. 2.3 bis 2.5) lassen sich weniger Textsorten bestimmen als vielmehr Domänen, also inhaltlich charakterisierte Textgruppen.[4] Wir werden daher auch im Folgenden, bezogen auf unsere quantitative Analyse, nicht von Textsorten sprechen, sondern von Textgruppen oder auch von Texttraditionen (verstanden als inhaltlich charakterisierte Textgruppen).

Als Basis der Untersuchung dienen Textausschnitte aus zwei Korpora, dem Referenzkorpus Frühneuhochdeutsch (ReF) und dem GerManC-Korpus (GerManC). Jeder Text wird als ein Vektor repräsentiert, der aus den Frequenzen seiner Ngramme gebildet wird. Mit Hilfe der Cosinus-Ähnlichkeit werden die Texte dann geclustert. Die grundliegende Idee ist, dass solche Cluster den traditionellen Textsorten zwar nicht entsprechen, aber doch sehr nahekommen können. Um die optimale Anzahl von Clustern zu bestimmen, werden die Cluster als Prädiktoren in einer Regressionsanalyse genutzt, die als abhängige Variable die relative Frequenz des *wh*-Relativpronomens (Lemma) vorhersagt. Das *wh*-Relativpronomen ist bekannt für seine „gattungsspezifische und stilistische Variation“ (Ebert 1986: 161), für seinen Anstieg in der Häufigkeit (sowie teils schon wieder Abnahme) in der frühen Neuzeit und eignet sich daher besonders gut für die Anwendung bei unseren ermittelten Textgruppen. Die Untersuchung zeigt, dass die statistische Analyse viele Erkenntnisse und/oder Annahmen aus der Literatur bestätigen kann: Gerade im Zeitraum vom 17. bis 18. Jahrhundert lassen sich inhaltlich klar charakterisierte Textgruppen erkennen, die mit der traditionellen Textsortengliederung übereinstimmen. Im früheren Zeitraum vom 14. bis 16. Jahrhundert sind die mit Hilfe von Expertenwissen vorgenommenen Textgliederungen hingegen, zumindest unserer Clusteranalyse zufolge, weniger klar voneinander abgegrenzt. Dies kann dahingehend interpretiert werden, als dass Faktoren wie die Einbettung in eine bestimmte historische Sinnwelt eine entscheidende Rolle bei der Ausbildung von Textsorten spielen (vgl. Schuster 2019: 229; zum Begriff der Sinnwelt siehe auch 2.1).

Der Aufsatz ist wie folgt gegliedert: In Abschnitt 2 gehen wir auf den Forschungshintergrund ein und behandeln Ansätze der klassischen Textsortengliederung (2.1) sowie Textsortenbetrachung im quantitativen Kontext (2.2). Im Anschluss daran werden die Korpora, auf der unsere quantitative Analyse beruht, beschrieben (2.3); zudem erläutern wir den theoretischen Hintergrund im Hinblick auf unsere Modellerstellung (2.4 zu Vektorraummodellen, 2.5 zur Clusterbildung mittels Regressionsanalyse). Der dritte Abschnitt behandelt die textsortenbedingte Verwendung des

4 Textsorten zeichnen sich vorwiegend durch Formeln oder bestimmte syntaktische Muster aus; in unserer quantitativen Analyse arbeiten wir aber vorwiegend mit Unigrammen, für Details zur Methode, s. 2.4 und 2.5.

wh-Relativpronomens in der frühen Neuzeit: Hier beschreiben wir die ermittelten Cluster (3.1), machen Angaben zur textsortenbedingten Verwendung des *wh*-Relativpronomens basierend auf der Forschungsliteratur und präsentieren und interpretieren schließlich die Vorhersagen unseres Regressionsmodells zum Auftreten des *wh*-Relativpronomens (3.3).

2 Forschungshintergrund

2.1 Texte, Textsorten und Textklassifikationen

Textsorten lassen sich definieren als „Teilmengen von Texten, die sich durch bestimmte relevante gemeinsame Merkmale beschreiben und von anderen Teilmengen abgrenzen lassen“ (Hartmann 1971: 22). Heinemann/Viehweger (1991: 148–169) bezeichnen typische Eigenschaften im Layout und im Bereich der Lexik/Grammatik als textinterne Merkmale, textexterne Eigenschaften umfassen die Funktion des Textes und den Kontext seiner Entstehung. Statt von Merkmalen kann man auch von Mustern der Textoberfläche und den kontextuellen Beschreibungsdimensionen sprechen (vgl. Bubenhofer 2009: 45 und Feilke 2003: 219). Darüber hinaus gibt es prototypische und weniger prototypische Vertreter einer Textsorte (vgl. Pfefferkorn 2005: 46–48). Textsorten zeichnen sich durch eine Musterhaftigkeit aus; diese Musterhaftigkeit steht immer in einem „Spannungsfeld von Konvention und Innovation“ (Koch 1997: 61) bzw. zwischen einer „Festlegung auf Verbindliches und Möglichkeit für Abweichungen zugleich“ (Fix 2008: 67). In ähnlicher Weise werden auch von Schuster/Haaf (2023) Textmuster verstanden, wenn sie schreiben, dass diese „an ihrer sprachlichen Musterhaftigkeit erkennbar“ (Schuster/Haaf 2023: 19) sind. Dementsprechend verstehen sie dann auch unter Textsortenwandel die „evolutionäre Umgestaltung von Textmustern“ (Schuster/Haaf 2023: 19). Bei der Herausbildung von Textsorten spielen Faktoren wie Geltungsgrad, kommunikative Reichweite und Einbettung in eine bestimmte historische Sinnwelt eine entscheidende Rolle (vgl. Schuster 2019: 229).

Es gibt unterschiedliche Herangehensweisen, wie Texte einer Textsorte zugeordnet werden können und eine Textklassifikation entsteht. Das frühneuhochdeutsche Lesebuch (Reichmann/Wegera 1988) basiert auf einer autorbezogenen Klassifikation. Hier steht der Autor mit seinen bestimmten zeitlichen, räumlichen, sozialen usw. Bindungen im Mittelpunkt. Reichmann (1996: 122–124) zufolge könnte man das Frühneuhochdeutsche aber auch unter textbezogenen Eigenschaften gliedern oder nach Kriterien einer leserbezogenen Klassifikation ordnen. Bei einer textbezogenen Klassifikation steht der Text mit seinen Eigenschaften im Mittel-

punkt (u. a. Lexik, Grammatik, Stil, Inhalt/Thema, Länge), bei einer leserbezogenen Klassifikation beruht diese auf dem Leser/der Rezeption des Textes.

Eine andere Herangehensweise an eine Textgliederung des Frühneuhochdeutschen findet sich bei Kästner et al. (2000): Die Autorinnen und Autoren verwenden keine differenzierende Textklassifikation, sondern streben ein „grobes Ordnungsraster auf sehr abstrakter Stufe [an], das für sehr große Textmengen einen ersten Zugriff erlaubt“ (Kästner et al. 2000: 1606). Ihr Raster basiert auf dem Konzept der „Sinnwelt“, in der Wirklichkeit interpretiert und in Texten mitgeteilt wird (vgl. Kästner et al. 2000: 1606). Die Autorinnen und Autoren gehen von fünf verschiedenen Sinnwelten aus (eine alltägliche, institutionelle, religiöse, wissenschaftliche und dichterische Welt), und jeder dieser Welten entsprechen spezifische Semantiken (vgl. Kästner et al. 2000: 1606–1607). Die institutionelle Welt zeichnet sich beispielsweise durch die „Regelung einzelner Ausschnitte des sozialen Lebens nach tradierten, oft genau definierten Begriffen und explizit geregelten Verfahrensnormen in den Bereichen Politik, Verwaltung, Recht und teilweise Wirtschaft“ (Kästner et al. 2000: 1606) aus. Mit den Institutionen sind bestimmte Stile und Formulierungsweisen verbunden wie der Kanzleistil (vgl. Kästner et al. 2000: 1606). Typische Textsorten dieser im 15. und 16. Jahrhundert wachsenden und sich ausdifferenzierenden Sinnwelt sind Gesetze, (Ver-)ordnungen, Verträge und Protokolle (vgl. Kästner et al. 2000: 1606–1607). Die Sichtweise auf Texte, wie sie bei Kästner et al. (2000) vorzufinden ist, entspricht in der Klassifikation nach Reichmann am ehesten dem textorientierten Zugang. Auch die quantitative Linguistik stellt den Text in den Mittelpunkt, wie im Folgenden zu sehen ist.

2.2 Textsortenbetrachtung und quantitative Linguistik/ Computerlinguistik

Im Kontext der quantitativen Linguistik haben sich zwei Herangehensweisen etabliert, die wir als induktiv und deduktiv bezeichnen. Ein induktiver Zugang basiert eine Textsortenzuordnung auf der philologischen Auseinandersetzung mit dem Kontext, in dem der Text entstanden ist; sprachliche Muster werden nach der Textsortenzuschreibung ermittelt (vgl. Schnelle 2020: 12). Ein deduktiver Ansatz hingegen stellt den Text selbst in den Mittelpunkt, die Auswahl der Merkmale trifft aber immer noch die Wissenschaftlerin/der Wissenschaftler. Ein Beispiel für die deduktive Herangehensweise ist die multidimensionale Analyse, wie sie von Biber/Conrad (2009: 225–230) für das (gegenwartssprachliche) Englisch vorgestellt wurde: Insgesamt 90 linguistische Strukturen (sowohl lexikalische als auch grammatische Merkmale) dienen ihm als Grundlage zur Klassifikation von Texten, bezogen auf ihre Funktion (Textsorte). Quantitative Textsortenbetrachtungen für das Deutsche finden

sich bei Bubenhofer/Scheurer (2014) und Bubenhofer/Spieß (2012). Diese Arbeiten sind eher korpusbasiert als korpusgetrieben, sie kombinieren quantitative mit qualitativen Elementen und berücksichtigen neben lexikalischen auch grammatische Elemente. Korpusgetriebene Arbeiten zum Deutschen finden sich für die Gegenwartssprache bei Scharloth (z. B. 2017, 2023) sowie für historische Texte bei Lasch (2023). Diese Ansätze basieren auf der Ermittlung von Keywords und Ngrammen.

Möglichkeiten zur Textsortenklassifikation sind auch in der Computerlinguistik bekannt: Bedingt durch das Aufkommen von Methoden, die von den Grundsätzen der distributionellen Semantik inspiriert sind, hat die korpusbasierte/-getriebene Modellierung semantischer Phänomene im letzten Jahrzehnt einen methodischen Paradigmenwechsel erfahren. Die Grundsätze der distributionellen Semantik fasst Firths (1957) Aphorismus „you shall know a word by the company it keeps“ zusammen (dies führte z. T. aber auch zu Missinterpretationen, vgl. Geeraerts 2017). Grundidee der distributionellen Semantik ist, dass Unterschiede/Ähnlichkeiten in der Bedeutung mit Unterschieden/Ähnlichkeiten in der Verteilung der sprachlichen Elemente korrelieren. Dieses Prinzip wurde in methodischer Hinsicht in der Form von Vektorraummodellen (oder auch neuronalen Einbettungsmodellen) umgesetzt. Solche Modelle repräsentieren den jüngsten Stand der Technik für alle Arten von Anwendungen im Bereich der Verarbeitung natürlicher Sprache (vgl. Pilehvar/Camacho-Collados 2020). Auch über die Computerlinguistik hinaus setzen sie sich allmählich in der theoretischen Linguistik durch (vgl. Lenci 2018, Boleda 2020, Lenci/Sahlgren 2023, Geeraerts et al. 2023). Der Anwendungsbereich reicht von der Identifikation von Polysemen und Synonymen hin zur Klassifikation von Themen in Texten. Letzteres tauchte zunächst im *Information Retrieval* auf: Bei einer Websuchanfrage müssen ähnliche und relevante Themen identifiziert und abgerufen werden und dafür benötigt man eine quantitative Operationalisierung (vgl. Clark 2015).

Bevor wir uns im Folgenden mit dem hier verwendeten Verfahren näher befassen, soll zunächst unsere Datengrundlage vorgestellt werden: Es sei daran erinnert, dass die quantitative Analyse und unsere Interpretation allein auf diesem Material erfolgt. Insofern kann unsere Herangehensweise auch als „bottom-up“ bezeichnet werden.

2.3 Beschreibung der verwendeten Korpora

Die beiden verwendeten Korpora[5] für den Zeitraum von 1350 bis 1800 setzen sich aus den beiden öffentlich zugänglichen Korpora „Referenzkorpus Frühneuhochdeutsch“ (Wegera et al. 2021) und „GerManC“ (Durrell et al. 2012) zusammen. Das „GerManC“ folgt dem Modell des „Bonner Frühneuhochdeutschkorpus“ in der zeitlichen Aufteilung (50-Jahres-Schritte) und in der Berücksichtigung verschiedener Schreiblandschaften. Das „Bonner Frühneuhochdeutschkorpus“ ist ein Vorläufer des heutigen „Referenzkorpus Frühneuhochdeutsch“ (vgl. Moser 2023: 465). Das „Referenzkorpus Frühneuhochdeutsch“ (im Folgenden: ReF) umfasst den Zeitraum von 1350 bis 1650,[6] das „GerManC“ den Zeitraum von 1650 bis 1800. Beide Korpora basieren auf einer Textsortengliederung, die auf einer philologischen Analyse beruht und den autorbezogenen Kriterien (s. 2.1) nahesteht. Darüber hinaus weisen beide Korpora ein breites Spektrum an verschiedenen Texten auf: Im ReF finden sich Rechts- und Geschäftstexte (RG), chronikalische und Berichtstexte (CB), Realientexte/Wissenschaftstexte (RE), unterhaltende Texte (UN), kirchlich-theologische Texte/Bibeln (KT) und erbauliche Texte (EB).[7] Im GerManC sind die Metadaten auf Englisch kodiert und lauten *drama, humanities, legal, narrative, newspaper, science, sermon*. Zum Teil finden sich in den Textsortenbezeichnungen Ähnlichkeiten zwi-

5 In einer ersten Version hatten wir diese beiden Korpora zu einem großen Korpus kompiliert. Leider sind die Unterschiede zwischen den beiden Korpora jedoch beträchtlich (z. B. unterschiedliches POS-Tagging, unterschiedliche Textlänge, verschiedene Textsorten) und die Ergebnisse der Clusteranalyse waren immer sehr stark von dem jeweiligen Korpus beeinflusst. Daher haben wir uns entschieden, die beiden Korpora getrennten Clusteranalysen zu unterziehen.
Unseres Wissens gibt es derzeit kein annotiertes und einheitliches Korpus, das den Zeitraum von 1350 bis 1800 abdeckt, und das wir daher als Grundlage für unsere quantitative Methode hätten nutzen können. Da wir unsere ermittelten Texttraditionen in Bezug auf das bekanntermaßen textsortenabhängige *wh*-Relativpronomen und dessen (vorhergesagte) Auftretenshäufigkeit testen möchten, sind wir auf ebendiesen Zeitraum angewiesen (das *wh*-Relativpronomen ist erstmals im 15. Jahrhundert belegt und hat seine Hochzeit vom 16. bis 18. Jahrhundert, vgl. 3.2).
6 Mit einer Ausnahme: Die Schrift *Gebrauch der Saurbrunnen* von Melchior Sebisch ist auf 1655 datiert. Im ReF ist die Schrift fälschlicherweise dem Zeitraum 1600–1650 zugeordnet; wir haben sie in der Darstellung in Tabelle 2 aus praktischen Gründen der Spalte „1600–1650“ (in der Kategorie „RE“) zugeordnet, nicht jedoch in unserer Analyse.
7 Unter RG fallen beispielsweise Rechtsquellen, Gerichts-, Dorf- und Polizeiordnungen, Berg- und Stadtrechte, Verwaltungstexte, Handels- und Zollakten. Zu CB gehören Berichte, Reisebücher und Geschichtsdarstellungen; Texte mit der Kennzeichnung „RE“ umfassen Chirurgien, Poetiken, Kochbücher, Rezeptbücher, Arzneibücher, Naturlehren, Astronomie- und Sprachbücher. Unter „UN“ fallen Volkslieder, Schauspiele, Romane und Volksbücher; zu „KT“ gehören theologische Fachprosa, Predigten, Legenden, geistliche Spiele, Flugschriften und Kirchenlieder. Texte mit der Kennzeichnung „EB“ umfassen Legenden, Trostbücher, Sterbebücher und Leichabdankungen.

schen den beiden Korpora (z. B. *legal* und RG sowie *narrative* und RE); für eine Clusteranalyse, die das ReF und GerManC zu einem Korpus kompilieren würde, müsste man aber auch die anderen Textsortenbezeichnungen über die beiden Korpora hinweg vereinheitlichen (also n*ewspaper* und CB, *science* und RE, *sermon* und EB). Schwierig würde die Zuordnung insbesondere von *sermon*, *humanities* und *drama*, da wir hierfür keine geeigneten bzw. möglichst übereinstimmenden Textsortenbezeichnungen im ReF vorfinden.

Ein Nachteil unserer Herangehensweise bzw. ein Nachteil, der sich aus der Wahl/Vorgabe durch die Korpora ergibt, besteht darin, dass wir nur Textausschnitte verwenden. Paratexte (z. B. Inhaltsverzeichnisse, Register, Vorworte) haben keinen Eingang in die Textauswahl gefunden. Auch weitere, für die philologische Analyse oder Textsortenzuordnung häufig wichtige Textausschnitte wie das Schlusskapitel stehen (wenn sie überhaupt im Korpus enthalten sind) „nur" gleichberechtigt neben anderen, aus philologischer Sicht möglicherweise weniger „wichtigen" Abschnitten (wie Abschnitte aus dem Mittelteil eines Textes). Auch die Bedeutsamkeit einzelner Texte kann unsere Herangehensweise nicht widerspiegeln: Im 15. und 16. Jahrhundert machten sich bspw. um die Einführung des Römischen Rechts leicht verständliche Rechtstexte wie der *Klagspiegel* des Conrad Heyden (um 1436), der *Laienspiegel* des Ulrich Tengler (ab 1509 im Druck) oder der *Rechtenspiegel* von Justin Göbler (1552 und später) sehr verdient, die zudem äußerst erfolgreich waren.[8] In unseren Korpora ist keiner dieser Rechtstexte vertreten. In anderen Worten spielen also Zufälligkeiten der Korpusbildung eine nicht unbedeutende Rolle: Unsere ermittelten inhaltlich charakterisierten Textgruppen basieren auf den Textausschnitten, die durch die Korpora vorgegeben sind.

Wir hätten sehr gerne die beiden Korpora, ReF und GerManC, zu einem Korpus kompiliert und auf Basis dessen inhaltlich charakterisierte Textgruppen quantitativ ermittelt und analysiert. Aufgrund zahlreicher Unterschiede zwischen den beiden Korpora mussten wir diese (ursprüngliche) Idee jedoch aufgeben:[9] Neben Unterschieden in der Textsortenklassifizierung finden sich auch Unterschiede zwischen den beiden Korpora bezogen auf den Umfang der verwendeten Texte: Die Textausschnitte des ReF haben ungefähr einen Umfang von 20 000 Wörtern (vgl. Herbers et al. 2021: 2–4) und sind damit deutlich länger als die Texte aus dem GerManC, die nur etwa 2.000 Wörter pro Textausschnitt umfassen (vgl. Durrell 2012: 1). Für eine quantitative Analyse sollten aber Texte gleicher Länge miteinander verglichen

8 Wir verdanken diesen Hinweis einer anonymen Gutachterin/einem anonymen Gutachter.

9 Unsere Versuche, mit einem kompilierten Korpus zu arbeiten, waren in der Praxis weniger erfolgreich, denn die ermittelten und analysierten Cluster waren trotz unserer Bemühungen, beide Korpora zu einem Korpus zusammenzuführen, immer noch klar in die beiden Korpora aufgeteilt. Siehe dazu auch Fußnote 5.

werden. Das ist in unserem Fall nur mit zwei getrennten Korpora möglich; auch würde es keinen (oder nur wenig) Sinn machen, das sowieso schon aus weniger Textausschnitten pro Textsorte bestehende ReF (im Vergleich zum GerManC) auf nur 2.000 Wörter pro Textausschnitt zu kürzen. Die Tabellen 1 und 2 geben einen Überblick über die Verteilung der verschiedenen Textsorten über die Zeiträume hinweg. Im Anhang finden sich zudem alle verwendeten Texte tabellarisch angeordnet unter Angabe des Zeitraums, Titels, Autors (sofern vorhanden), der aufgrund philologischer Analyse zugeordneten Textsorte, der Anzahl der Wörter pro Textausschnitt sowie der Anzahl der Relativsatzmarker pro Textausschnitt. Gesamt wurden 463 Textausschnitte verwendet, im ReF sind das 148 Textausschnitte, im GerManC 315 Textausschnitte. In Wörtern bzw. Tokens ausgedrückt umfasst das ReF ca. 2,96 Millionen Wörter, das GerManC ca. 630 000 Wörter; zusammen ergibt das also knapp 3,6 Millionen Wörter, auf der unsere Analyse beruht.

Tab. 1: Verteilung und Anzahl der Textsorten nach Zeiträumen im ReF

	1350–1400	1400–1450	1450–1500	1500–1550	1550–1600	1600–1650
RG	4	3	3	8	3	0
CB	0	1	8	4	2	3
RE	1	1	6	12	7	3
UN	2	2	11	5	4	2
EB	5	4	7	6	4	2
KT	4	4	3	11	2	1

Tab. 2: Verteilung und Anzahl der Textsorten nach Zeiträumen im GerManC

	1650–1700	1700–1750	1750–1800
legal	15	15	15
newspaper	15	15	15
science	15	15	15
narrative	15	15	15
drama	15	15	15
sermon	15	15	15
humanities	15	15	15

Ein weiterer Unterschied zwischen den beiden Korpora betrifft das Medium, also Handschrift vs. Druck. GerManC arbeitet nur mit Drucken (vgl. Durrell 2012: 1), im ReF finden sich sowohl Handschriften als auch Drucke: In den Zeiträumen 1350–1400 und 1400–1450 wurden nur Handschriften, in den Zeiträumen 1450–1500 und 1500–1550 wurden sowohl Handschriften als auch Drucke verwendet (vgl. Herbers et al. 2021: 3). Tabelle 3 gibt einen Überblick über die Anzahl der verwendeten Handschriften und Drucke für den Zeitraum 1450 bis 1550.

Tab. 3: Anzahl der Handschriften und Drucke für den Zeitraum 1450 bis 1550

		1450–1500	1500–1550
RG	Handschrift	0	6
	Druck	3	2
CB	Handschrift	3	3
	Druck	5	3
RE	Handschrift	2	6
	Druck	4	6
UN	Handschrift	8	2
	Druck	3	3
EB	Handschrift	7	1
	Druck	0	5
KT	Handschrift	1	2
	Druck	2	9

Die Unterscheidung in Handschrift vs. Druck ist für uns relevant, da Drucke üblicherweise einen höheren Grad an Standardisierung aufweisen als Handschriften. Eigenheiten in der Schreibung, die im ReF eventuell auf Unterschiede im Medium zurückzuführen sind, haben wir dadurch ausgeschlossen, dass wir als Basis der Vektorraummodelle Lemmata verwendet haben. Die Lemmatisierung erfolgte im ReF nach dem Deutschen Wörterbuch von Jacob und Wilhelm Grimm (s. DWB) (vgl. Herbers et al. 2021: 24). Darüber hinaus ist die Verwendung von Lemmata im ReF sinnvoll, um regionalsprachliche Besonderheiten, die sich in der Lexik/Orthographie widerspiegeln, auszuschließen: Die (Kanzlei)Schriftlichkeit in der Volkssprache ist nämlich, so Tschirch (1989: 97), noch bis in das 16. Jahrhundert stark lokal und regional orientiert. Meier (2012: 10) stellt ebenfalls fest, dass sich bereits zu Beginn des

16. Jahrhunderts die Kanzleisprachen aus unterschiedlichen Regionen sehr nahestehen. Allerdings machen sich regionalsprachliche Einflüsse nicht nur in kanzleisprachliche Texten, sondern generell in allen Texten bis Ende des 16. Jahrhunderts im unterschiedlichen Ausmaß bemerkbar: Das wird beispielsweise gut sichtbar in der frühneuhochdeutschen Grammatik (Ebert et al. 1993), in der explizit auf das Fehlen einer Prestigevarietät im Frühneuhochdeutschen hingewiesen wird (vgl. Ebert et al. 1993: 8) und in der dementsprechend ein Raumraster mit den unterschiedlichen Schreibregionen den „sachlichen und terminologischen Rahmen der Grammatik" (Ebert et al. 1993: 6) bildet. Regionalsprachliche Einflüsse werden insbesondere dann sichtbar, wenn nicht nur die Schriftlichkeit in der Nachfolge Martin Luthers im Vordergrund steht, sondern auch die des süddeutschen Raums (vgl. z. B. Mattheier 2003: 215–218 als Überblick sowie Wiesinger 2012 für den bairisch-habsburgischen Raum).

Sowohl im ReF als auch im GerManC haben wir eine Stoppwortliste verwendet, damit die Vektorraummodelle nicht auf Basis von Funktionswörtern gebildet werden. Wir haben unterschiedliche Stoppwortlisten ausprobiert: Für das GerManC Stoppwortlisten mit den 50 oder mit den 100 häufigsten Wortformen, für das ReF Stoppwortlisten mit den 50 oder 100 häufigsten Lemmata. Zudem haben wir eine von den Korpora unabhängige Stoppwortliste ausprobiert, die über das NLTK (Natural Language Toolkit) (vgl. Bird et al. 2009) verfügbar ist. Dabei stellte sich heraus, dass für das ReF eine Stoppwortliste mit den 50 häufigsten Lemmata das beste Ergebnis lieferte, für das GerManC eine Stoppwortliste von 100 Wortformen (für Details, s. 2.4 und 2.5). Da das GerManC nur Drucke beinhaltet und zudem den Zeitraum von 1650–1800 umfasst, können wir davon ausgehen, dass die Texte einen sehr hohen Grad an Standardisierung aufweisen. Wir arbeiten im GerManC daher nicht mit Lemmata, sondern mit Wortformen. Es sind vor allem zwei Gründe, warum wir im GerManC nicht mit Lemmata gearbeitet haben bzw. nicht damit arbeiten mussten. Zum einen waren Lemmata im ReF notwendig, um die zahlreichen Unterschiede in der Lexik/Orthographie zu neutralisieren. Das GerManC hingegen startet um 1650 und geht bis 1800 – Unterschiede in der Lexik/Orthographie sind hier aufgrund der Standardisierung weitestgehend auszuschließen. Zum anderen sind wir auf der Wortebene geblieben, da es auf dieser Ebene einfacher ist, klare Mehrwortausdrücke (Bi- und Trigramme) bzw. feste Formeln oder Textmuster zu identifizieren (für Details s. 2.4 und 2.5).

2.4 Vektorraummodelle

Grundlegendes Prinzip bei VRM ist die Tatsache, dass der Inhalt von Texten als Vektor operationalisiert wird, also eine geordnete Folge von Zahlen in einigen Dimensionen repräsentiert wird. Die Texte werden in dieser Form semantisch

Tab. 4: Beispiel einer Text-Begriffs-Matrix mit Angabe der Häufigkeit von Begriffen (Unigramme) in vier Texten

	tochter	*mein*	*koenig*	*ihr*	*gott*	*jesu*
doc1	284	89	125	134	13	3
doc2	163	178	114	92	23	0
doc3	50	4	5	7	123	184
doc4	22	2	0	1	160	354

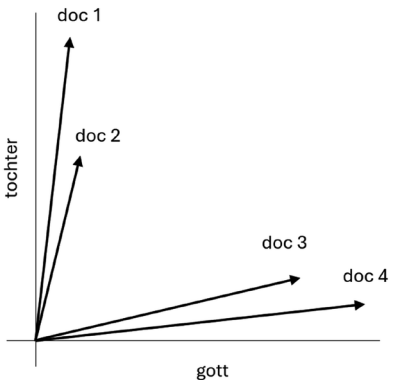


Abb. 1: Beispiel eines Vektorraums mit zwei Dimensionen *(tochter* und *gott*) und 4 Texten

repräsentiert. Die semantischen Vektoren kann man sich bildlich so vorstellen, dass ihre Zahlen Koordinaten in einem geometrischen Raum darstellen. Dadurch wird das Dokument, das sie repräsentieren, in einem solchen Raum verortet. Ein direkter Vorteil einer vektoriellen und räumlichen Darstellung von Themen und Bedeutungen besteht darin, dass die Verwandtschaft zwischen Texten als ein kontinuierliches Maß behandelt wird (indem der Abstand zwischen Vektoren berechnet wird).

Die numerischen Informationen, aus denen sich der Vektor eines bestimmten Textes zusammensetzt, werden gewonnen, indem zunächst das Dokument gelesen wird. Dabei werden alle Wörter, oder besser gesagt, Merkmale, die für die Erfassung des Inhalts als relevant erachtet werden, abgerufen. Jedes einzelne Merkmal stellt eine Dimension des Vektors dar. Für jedes dieser Merkmale wird seine Häufigkeit innerhalb des Textes registriert. Führt man dieses Verfahren für viele verschiedene Texte durch, so erhält man eine Text-Merkmals-Matrix mit Häufigkeitsdaten: Die Zeilen stellen die Texte dar, die Spalten die Merkmale und die Zellen die Frequenzen. Ein Beispiel für eine solche Matrix ist in Tabelle 4 dargestellt: Vier fiktive Dokumente erhalten jeweils einen Vektor mit Angabe der Frequenz; doc1 und doc2

haben ähnliche Frequenzprofile: Begriffe wie *tochter, mein, koenig* und *ihr* treten sehr häufig auf und Begriffe wie *gott* und *jesu* kommen nur gelegentlich vor. Da die Merkmale in diesem Beispiel Begriffe sind, kann man auch von einer Text-Begriffs-Matrix (anstatt einer Text-Merkmals-Matrix) sprechen. Da es mühsam und wenig informativ ist, diese Vektoren auf der Suche nach Ähnlichkeiten zu betrachten, besteht der letzte Schritt in der Berechnung eines Abstandsindex zwischen allen Vektorpaaren. Der Abstandsindex erfasst die semantische Ähnlichkeit zweier Texte. Abbildung 1 zeigt die geometrische Interpretation (eines Teils) der fiktiven Matrix in Tabelle 4, basierend auf den Dimensionen *tochter* und *gott*. Die Pfeile entsprechen den Vektoren, und je kleiner der Winkel zwischen den Vektoren ist, desto größer ist die Ähnlichkeit zwischen den Texten: So ähneln sich doc1 und doc2 insofern, als dass sie eine hohe Frequenz für *tochter* und eine niedrige Frequenz für *gott* aufweisen; bei doc3 und doc4 sind die Häufigkeitszahlen genau umgekehrt.

In Tabelle 5 werden die Vektorabstände zwischen den verschiedenen Texten auf Grundlage der in der Text-Begriffs-Matrix erfassten Häufigkeiten in Tabelle 1 dargestellt. Diese Abstände werden mithilfe der Cosinusfunktion (Cosinus-Abstandsformel) berechnet, wobei höhere Werte größere Abstände anzeigen.

Tab. 5: Beispiel einer Distanzmatrix

	doc1	**doc2**	**doc3**	**doc4**
doc1	0	0.102	0.770	0.928
doc2	0.102	0	0.799	0.930
doc3	0.770	0.799	0	0.028
doc4	0.928	0.930	0.028	0

Die Konstruktion von Vektorraummodellen beinhaltet die Wahl einer optimalen Kombination von Werten für die zahlreichen Hyperparameter, um eine angemessene Modellierung semantischer Phänomene zu ermöglichen. Die Forschung hat gezeigt, dass die Festlegung solcher Hyperparameterwerte von vielen Faktoren abhängen kann, z. B. von der Art des Korpus, von der Forschungsfrage, in die die Modelle integriert werden sollen, und auch von der semantischen Struktur der einzelnen Wörter (vgl. Geeraerts et al. 2023). Anders formuliert bedeutet das, dass man gar nicht wissen kann (bzw. es wäre ein entmutigendes Unterfangen), welche ideale Kombination von Hyperparametern die besten Ergebnisse liefern wird. Die in dieser Arbeit verwendete Strategie kann daher als agnostisch beschrieben werden, denn sie beruht auf dem Ausprobieren einer ganzen Reihe von Werten für jeden Hyperparameter, um eine große Anzahl von Modellen zur Auswahl zu haben.

Für unsere Modelle haben wir die Hyperparameter Kontextmerkmale, minimale Frequenzgrenze für Kontextmerkmale, SVD für dichtere Vektoren, maximale Texthäufigkeit für Kontextmerkmale und Stoppwortliste mit unterschiedlichen Werten variiert (s. Tabelle 6). Das bedeutet, dass wir für jedes Korpus, ReF und GerManC, 4*4*2*3*3 = 288 Vektorraummodelle und die daraus resultierende Cosinusdistanzmatrix erstellt haben, die als Input für die Regressionsanalyse verwendet wurde (für Details s. 2.5).

Tab. 6: Hyperparameter und Werte für die Erstellung der Vektorraummodelle

Hyperparameter	**Werte**
Kontextmerkmale	Unigramme/ Bigramme/ Unigramme+Bigramme/ Bigramme+Trigramme
Minimale Frequenzgrenze für Kontextmerkmale	1/ 5/ 10/ 20
Singular Value Decomposition (SVD) für dichtere Vektoren	SVD/ ohne SVD
Maximale Texthäufigkeit für Kontextmerkmale (im Verhältnis zu allen Texten, d. h. 100 % = alle Texte)	100 %/ 80 %/ 70 %
Stoppwortliste	50 häufigsten Frequenzen/ 100 häufigsten Frequenzen/ NLTK-basierte Liste

Sobald diese letzte Matrix mit den Vektorabständen zwischen den einzelnen Texten gebildet wurde, kann sie durch Anwendung einer Dimensionalitätsreduktionstechnik wie t-SNE (vgl. Van der Maaten/Hinton 2008) visualisiert werden (t-SNE steht für *t-distributed stochastic neighbor embedding*). Mithilfe dieser Technik wird die Dimensionalität des Textraums reduziert, um ihn interpretierbar zu machen.[10] Die verbliebenen zwei Dimensionen sind ganz neu erzeugte Dimensionen, die sich (als Dimension) oft nicht mehr sinnvoll interpretieren lassen. Allerdings ermöglichen die neu erzeugten Dimensionen eine angenehme visuelle Darstellung der signifikanten Muster zwischen allen Texten.

Die Abbildungen 2 und 3 zeigen den reduzierten Vektorraum aller Texte aus den beiden Korpora. Eine solche Visualisierung bietet den Vorteil, dass zusätzliche Informationen in Form von Farb- und Formkodierungen projiziert werden können. Diese zusätzlichen Informationen sind mit den Texten durch die Metadaten der

10 Ein solcher Raum besteht aus so vielen Dimensionen, wie es Texte gibt. Er ist daher unüberschaubar groß.

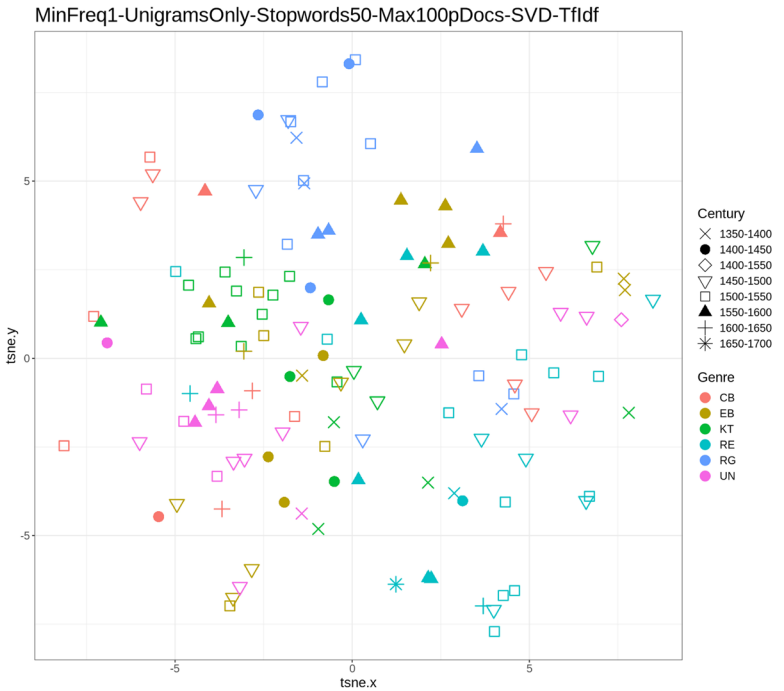


Abb. 2: t-SNE-Diagramm der Ähnlichkeitsmatrix der Texte aus ReF (Farbcodierung: Textsorten, Formcodierung: Jahrhunderte)

beiden Korpora verbunden: In den Abbildungen 2 und 3 zeigen verschiedene Formen verschiedene Jahrhunderte an, während verschiedene Farben die unterschiedlichen Textsorten darstellen. Insbesondere bei Abbildung 3 mit den Texten aus dem GerManC gibt es zahlreiche Übereinstimmungen zwischen der semantischen Vektormodellierung und den gegebenen Textsortenzuordnungen: Es sind Gruppen von Texten zu erkennen, die Texte derselben Textsorte enthalten. Dies bedeutet, dass die textbasierte Vektorraummodellierung in der Lage ist, Korpustexte so zu gruppieren, wie sie auch auf Basis sorgfältiger philologischer Arbeit möglich ist. Bis auf *humanities*, das im Vektorraum verteilt ist, sind die einzelnen Textsorten gut als solche zu erkennen, auch wenn es Unterschiede in der Dichte gibt wie bspw. bei *science* und *newspaper*. In Abbildung 2 sind ebenfalls Textgruppierungen zu erkennen, allerdings weniger deutlich: Gut erkennbar als Gruppierungen sind RE unten rechts ebenso wie RG oben in der Mitte. Links im unteren Teil findet sich noch eine kleine Ansammlung von UN, wobei diese Gruppierung stärker im Raum verstreut ist als RE und RG, denn UN streut bis weit nach rechts. Abgesehen davon trägt die vorgegebene Textsortenklassifikation nicht zur Strukturierung des Raumes bei: Die Texte mit unterschiedlichen Farben sind untereinander verstreut.

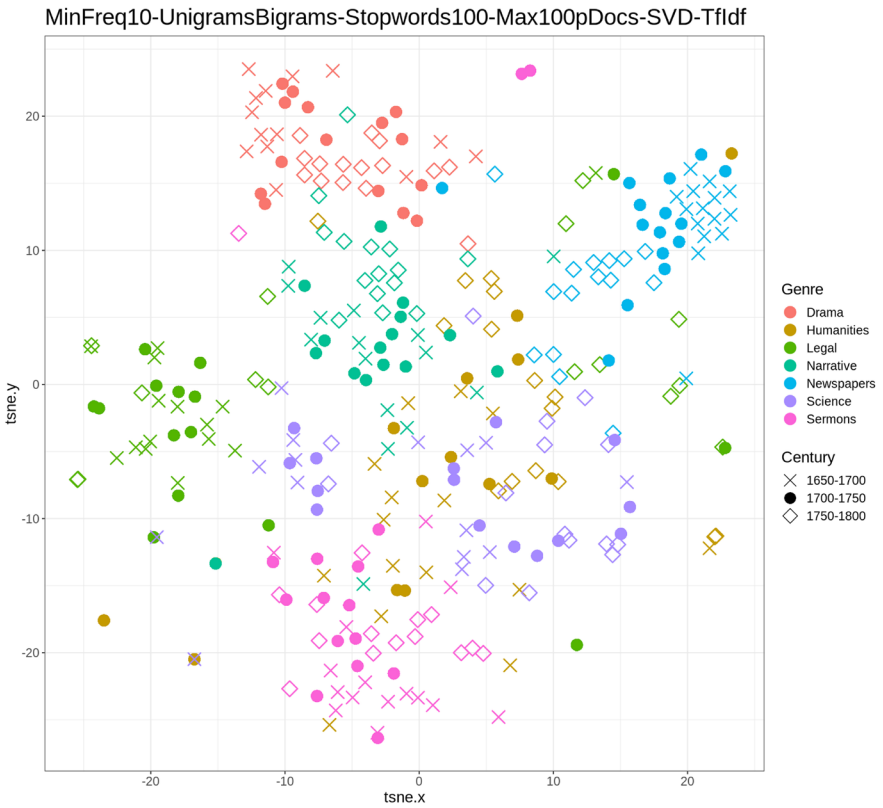


Abb. 3: t-SNE-Diagramm der Ähnlichkeitsmatrix der Texte aus GerManC (Farbcodierung: Textsorten, Formcodierung: Jahrhunderte)

Für die beobachteten Muster kann es mehrere Erklärungen geben. Die erste ist deskriptiver Natur: Die Schwierigkeit, die Texte in den ersten Jahrhunderten auseinanderzuhalten und ihre klare Gruppierung in den letzten Jahrhunderten spiegeln eine zunehmende Spezialisierung und Konventionalisierung der einzelnen Textsorten im Laufe der Zeit wider. Eine explorative Analyse wie diese hier könnte die Hypothese aufstellen, dass viele Wörter und Themen erst in späteren Jahrhunderten für bestimmte Textsorten indexikalisch werden. Die zweite Überlegung ist methodischer Natur: Die Texte aus Abbildung 3 stammen aus einem anderen Korpus als die Texte aus Abbildung 2, so dass Gründe, die mit der Zusammenstellung des Korpus zusammenhängen, in den Vordergrund treten: Hierbei zu nennen wären beispielsweise Unterschiede im Medium (Handschrift vs. Druck) und ein längerer Zeitraum von 3 Jahrhunderten im ReF (vs. 1,5 Jahrhunderten im GerManC). Da wir im ReF Lemmata genutzt haben, können Unterschiede zwischen den beiden

Darstellungen aufgrund geringerer Einheitlichkeit in der Lexik/Orthographie ausgeschlossen werden.

2.5 Modellerstellung: Clustering auf Grundlage eine Regressionsanalyse

Die ursprüngliche Ähnlichkeitsmatrix der Texte wird nun einer Clusteranalyse unterzogen, um Gruppierungen von Texten auf der Grundlage ihrer Ähnlichkeit zu erfassen. Mit anderen Worten: Was im oberen Abschnitt informell als Gruppierungen erkannt wurde, kann durch eine solche Technik robuster generiert werden. Ein wichtiger Aspekt bei der Anwendung von Clustertechniken ist die Bestimmung der Anzahl von Clustern, in die man den Textraum aufteilen möchte. In dieser Studie werden wir einen Ansatz verwenden, bei dem die optimale Clustering-Lösung diejenige ist, die die beste Vorhersagekraft der untersuchten linguistischen Variable (in unserem Fall das *wh*-Relativpronomen) aufweist. Wir ermitteln die optimale Clustering-Anzahl mithilfe einer Regressionsanalyse: Während die Modellierung von Texttraditionen mithilfe von Vektorraummodellen ein bekanntes Verfahren ist, ist hingegen die Verwendung einer Regressionsanalyse für eine bestimmte sprachliche Variable, zur Ermittlung der Texttraditionen/Cluster, neu und wird hier erstmals verwendet.

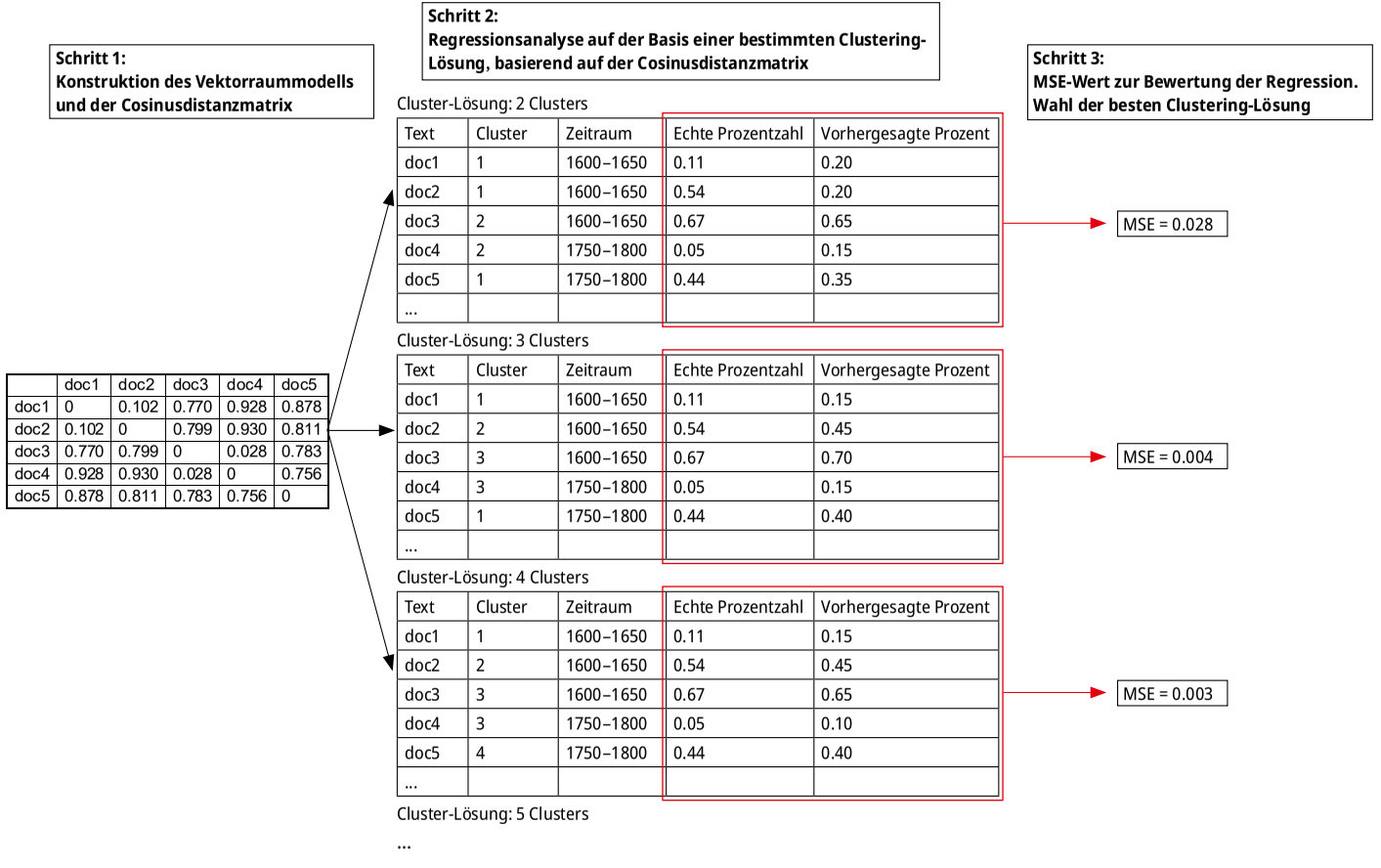
Für die Regressionsanalyse und das Clustering benötigen wir Angaben zu den *wh*-Relativpronomen: Die Datenstichprobe für das ReF umfasst 1.707 *wh*-Relativpronomen sowie 5.496 weitere Relativsatzmarker (davon 5.199 *d*-Relativpronomen). Im GerManC finden sich 2.678 *wh*-Relativpronomen sowie 5.345 weitere Relativsatzmarker (davon 3.686 *d*-Relativpronomen). Die Belege wurden in beiden Korpora anhand des POS-Tags extrahiert.[11] Es sei darauf hingewiesen, dass wir mit dem Lemma des *wh*-Relativpronomens arbeiten, d. h. das Regressionsmodell sagt die Wahrscheinlichkeit irgendeiner Form des *wh*-Relativpronomens voraus und nicht die Wahrscheinlichkeit von *welcher* vs. *welche* vs. usw.. Die Anzahl der *wh*-Relativsatzmarker spielt in unserem Modell eine doppelte Rolle: Zum einen wird sie genutzt, um mithilfe einer Regressionsanalyse Cluster zu bilden (s. 3.1 zur Beschreibung der ermittelten Cluster). Zum anderen werden aber auch die Vorhersagen des

11 Weitere Relativsatzmarker, die in beiden Korpora automatisch extrahiert werden konnten, sind das Relativpronomen (mit seinen flektierten Formen) *der/die/das*, die Relativpartikel *so*, der Relativsatzmarker *was* sowie relativische Pronominal- bzw. Präpositionaladverbien in Distanzstellung wie *da […] in* und nicht in Distanzstellung wie *daran, darüber*. Verwendete Befehle: PAVREL, PTKREL, DRELS, PAVRELAP, DWS, AVREL (im „Referenzkorpus Frühneuhochdeutsch“) sowie PTKREL, PAVREL, PWAVREL, PWREL im GerManC. Vgl. Herbers et al. (2021) und Durrell et al. (2012) zur Aufschlüsselung der Suchbefehle und zu ihrer Dokumentation.

Regressionsmodells zu Häufigkeiten in den Clustern analysiert und interpretiert (s. 3.3 zur Interpretation der Vorhersagen). Die Distribution des linguistischen Phänomens (*wh*-Relativpronomen) in unseren Korpora spielt bei unserem methodischen Vorgehen eine entscheidende Rolle, da diese Verteilung (über alle Texte hinweg) als Maßstab für die Auswahl der am besten geeigneten Clustering-Lösung verwendet wird. Unter allen Regressionsanalysen, die als Input-Prädiktor eine bestimmte Clustering-Lösung eines bestimmten Vektorraummodells haben, wählen wir die eine Regressionsanalyse mit dem geringsten Testfehler aus (d. h. die Analyse, die die Verteilung des *wh*-Relativpronomens über Cluster und Zeiträume am besten modelliert). Sobald eine solche Regressionsanalyse identifiziert wurde, wird sie genauer untersucht, um zu beurteilen, warum diese spezifische Clusterlösung bei der Modellierung der Häufigkeitsveränderung des *wh*-Relativpronomens über die Zeit und die Textsorten hinweg am erfolgreichsten zu sein scheint. Um eine gewisse Zirkularität zu vermeiden, die entstehen könnte, wenn man das Clustering auf der Grundlage von Vektorraummodellen mit *wh*-Relativpronomen als Kontextmerkmalen bewerten würde, haben wir beschlossen, zunächst alle *wh*-Relativpronomen aus den Texten zu entfernen, bevor wir die Vektorraummodelle erstellen.

In der Analyse fungieren Cluster und Zeitraum (in 50-Jahres-Schritten) als zwei kategoriale Prädikatoren (unabhängige Variablen) und die Angabe der Häufigkeit des *wh*-Relativpronomens pro Text ist die numerische abhängige Variable. Wir haben uns primär aus Gründen der Modellkomplexität dafür entschieden, „Periode“ als kategorialen Prädiktor beizubehalten und nicht als metrische Variable. Darüber hinaus ist es auch im Hinblick auf die Zuverlässigkeit/Belastbarkeit besser, zeitbezogene Prädiktoren als kategoriale Daten und nicht als kontinuierliche Daten zu betrachten. Die Modellspezifikation der LASSO-Regression in R erfordert als Eingabe die Interaktionsprädiktoren als One-Hot-Vektoren und nicht als zwei getrennte Prädikatoren und ihre Interaktion (wie es bei der linearen Regression mit Im oder Imer üblich ist). Um herauszufinden, welche Clusterlösung auf der Grundlage einer Regressionsanalyse mit der Prozentzahl der *wh*-Relativsatzmarker als abhängige Variable am besten zu den Daten passt, ist das folgende iterative Verfahren erforderlich: Jede Iteration ist eine Regressionsanalyse mit einer Clusterlösung als Prädiktor, die ein Cluster mehr hat als die vorherige Iteration. Eine „Clusterlösung als Prädiktor“ bedeutet, dass jeder Text ein Cluster-Label erhält und dass ein bestimmter Prozentsatz der *wh*-Relativpronomen auf dieses Label regressiert wird. Dieser Clusterprädiktor interagiert dann mit dem Zeitraum als Prädiktor, so dass eine einzelne Prädiktor-Ebene auf eine Kombination eines bestimmten Clusters mit einer bestimmten Zeitperiode hinausläuft: Ein Text aus dem GerManC erhält so ein Prädiktor-Label namens „cluster18_1650–1700“.

Wir beginnen beispielsweise damit, den Textraum in 2 Cluster zu unterteilen, und dies in einer Interaktion mit 7 Zeitabschnitten für das ReF und 3 Zeitabschnit-



	doc1	doc2	doc3	doc4	doc5
doc1	0	0.102	0.770	0.928	0.878
doc2	0.102	0	0.799	0.930	0.811
doc3	0.770	0.799	0	0.028	0.783
doc4	0.928	0.930	0.028	0	0.756
doc5	0.878	0.811	0.783	0.756	0

Text	Cluster	Zeitraum	Echte Prozentzahl	Vorhergesagte Prozent
doc1	1	1600–1650	0.11	0.20
doc2	1	1600–1650	0.54	0.20
doc3	2	1600–1650	0.67	0.65
doc4	2	1750–1800	0.05	0.15
doc5	1	1750–1800	0.44	0.35
...				

Text	Cluster	Zeitraum	Echte Prozentzahl	Vorhergesagte Prozent
doc1	1	1600–1650	0.11	0.15
doc2	2	1600–1650	0.54	0.45
doc3	3	1600–1650	0.67	0.70
doc4	3	1750–1800	0.05	0.15
doc5	1	1750–1800	0.44	0.40
...				

Text	Cluster	Zeitraum	Echte Prozentzahl	Vorhergesagte Prozent
doc1	1	1600–1650	0.11	0.15
doc2	2	1600–1650	0.54	0.45
doc3	3	1600–1650	0.67	0.65
doc4	3	1750–1800	0.05	0.10
doc5	4	1750–1800	0.44	0.40
...				

Abb. 4: Überblick zu den einzelnen Schritten bei der Modellerstellung

ten für das GerManC (s. zu den einzelnen Schritten Abbildungen 4). Wir verwenden diese 2*X oder 2*Y Cluster als Prädiktorwerte für die Regression auf den Prozentsatz der *wh*-Relativpronomen (s. Schritt 2 in Abbildung 4). Die verwendete Clustertechnik heißt *hierachical agglomerative clustering with Ward linkage function.* Schließlich leiten wir ein Maß für die Anpassungsgüte der Regression ab (s. Schritt 3 in Abbildung 4). In der nächsten Iteration unterteilen wir denselben Textraum in 3 Cluster und verwenden diese 3 Cluster als Prädiktoren in Interaktion mit dem Zeitraum und so weiter, bis eine angemessene Anzahl von Clusterlösungen getestet wurde (in unserem Fall haben wir bei 100 Clustern aufgehört). Das Ergebnis dieses Verfahrens ist eine Liste von MSE-Werten pro Regressionsmodell mit einer bestimmten Clusterlösung (s. Schritt 3 in Abbildung 4). Der spezifische Anpassungsgüte-Index, der die Qualität des Regressionsmodells misst, variiert ebenfalls je nach Art der abhängigen Variable. Bei unserer Variablen entscheiden wir uns für den mittleren quadratischen Fehler (MSE: *mean squared error*) auf die Testmenge. Das genaue Regressionsmodell, das in dieser Studie verwendet wurde, ist eine LASSO-Regression mit k-facher Kreuzvalidierung, die besonders gut geeignet ist, wenn man viele verschiedene Prädiktoren hat, wie in unserem Fall den Interaktion-Prädiktor Cluster x Zeitraum.[12] Bei LASSO werden die Koeffizienten irrelevanter Ebenen auf null reduziert, wodurch sie effektiv entfernt werden und zu einem einfacheren und besser interpretierbaren Regressionsmodell führen (s. James et al. 2021: 241–251, 264–267; für Anwendungen in der Linguistik s. Van de Velde/Pijpops 2021). Das Endergebnis dieser iterativen Verfahren ist eine Liste von MSE-Werten pro Regressionsmodell mit spezifischer Cluster-Lösung. Da wir 288 Vektorraummodelle (4 × 4 × 3 × 3 × 2) konstruiert haben, wurden 28800 Clustering-Lösungen bewertet.

In den Abbildungen 5 und 6 ist die Kurve der MSE-Werte auf der y-Achse im Verhältnis zu der zunehmenden Anzahl von Clustern (mit einem Maximum von 100 Clustern) auf der x-Achse zu sehen. Je niedriger dieser Wert ist, desto besser kann das Regressionsmodell mit dieser spezifischen Clusterlösung die Verteilung der Prozentsätze der *wh*-Relativpronomen vorhersagen. Die beiden Abbildungen zeigen nicht den Verlauf der MSE-Werte für einzelne Modelle, sondern stellen den Durchschnitt über vier Gruppen von Text-Vektorraummodellen dar, d. h. sie bilden nur die aggregierten Werte von verschiedenen Gruppen von Ngrammen ab: solche, die mit Bigramm-Merkmalen erstellen wurden (rote Linie); solche, die mit Bigramm- und Trigramm-Merkmalen kombiniert erstellt wurden (blaue Linie); solche, die mit Unigramm- und Bigramm-Merkmalen erstellt wurden (grüne Linie); und solche, die

12 Die formale Spezifikation des Regressionsmodells ist der Standardfehler (*standard error*), berechnet mit der Funktion geom_smooth in ggplot2 in R. Die genaue Methode der Glättung ist LOESS (Local Polynomial Regression Fitting).

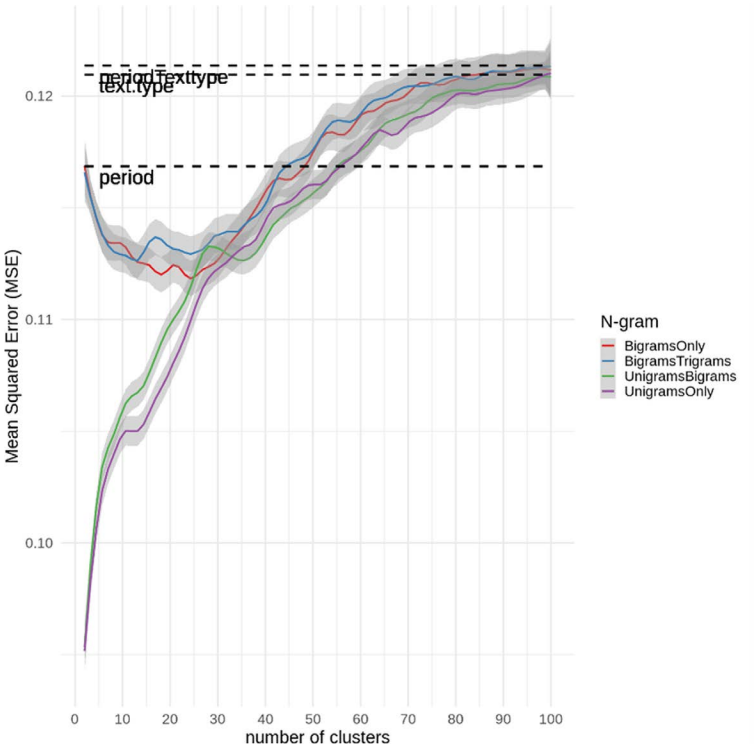


Abb. 5: MSE-Werte (Mittelwerte) für 4 Gruppen von Vektorraummodellen (farbige Linien) für das ReF (Standardfehler grau schattiert)

nur mit Unigramm-Merkmalen erstellt wurden (in lila). Die graue Schattierung zeigt jeweils den Standardfehler, stellt also die Genauigkeit der ermittelten Mittelwerte dar. Die beiden Abbildungen zeigen also, welche Hyperparameter (hier dargestellt am Beispiel der Kontextmerkmale) für die Konstruktion der semantischen Räume mit unseren Daten am besten funktionieren. In Abbildung 5, die die MSE-Werte für das ReF darstellt, ist beispielsweise deutlich zu erkennen, dass Vektorraummodelle, die ausschließlich Unigramm-Merkmale verwenden, nur wenige Cluster benötigen, um die niedrigsten und somit optimalen MSE-Werte zu erreichen. Im Gegensatz dazu erreichen Modelle mit Bigramm-Merkmalen ihre niedrigsten MSE-Werte im Durchschnitt bei 20 Clustern. Die Diagramme zeigen für beide Korpora, dass semantische Räume, die mit Unigramm- oder Unigramm+Bigramm-Merkmalen generiert wurden, signifikant besser sind als solche mit Bigramm- oder Bigramm+Trigramm-Merkmalen. Dies deutet darauf hin, dass Mehrworteinheiten oder Formeln nicht so nützlich sind wie bloße Unigramm-Merkmale, um die Zu- und Abnahme von *wh*-Relativpronomen zu erklären.

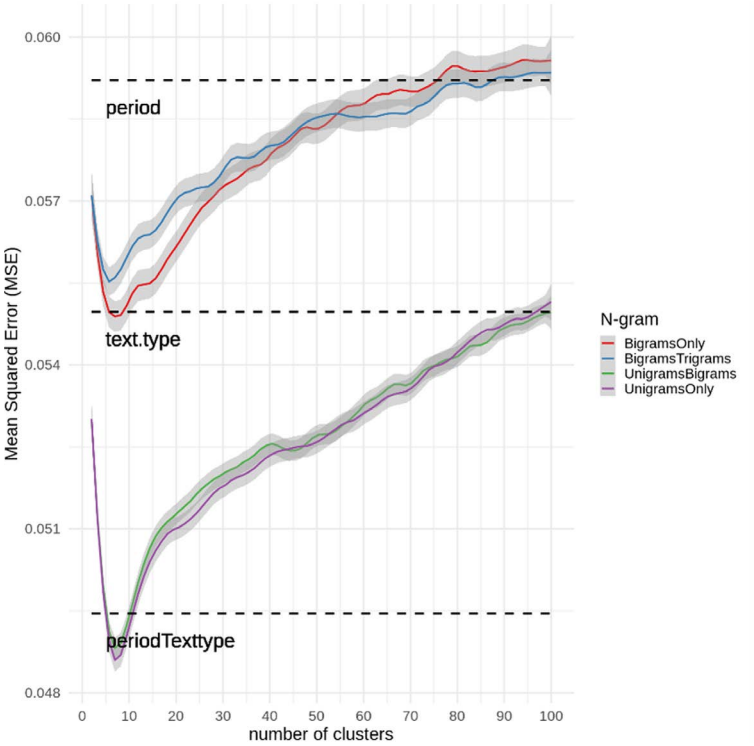


Abb. 6: MSE-Werte (Mittelwerte) für 4 Gruppen von Vektorraummodellen (farbige Linien) für das GerManC (Standardfehler grau schattiert)

Um das beste Regressionsmodell zu ermitteln, müssen wir nun prüfen, welches Modell den niedrigsten MSE-Wert über alle semantischen Räume und alle Clustering-Lösungen hinweg erreicht (man spricht auch vom globalen Minimum). Die MSE-Werte für ein einzelnes Modell (und ein Cluster innerhalb des Modells) erhält man, indem man eine Tabelle mit allen Clusterlösungen von allen Modellen sortiert. Die Abbildungen 5 und 6 können uns dies also nicht verraten: Diese stellen nämlich nicht die MSE-Werte für ein einzelnes Modell dar und sollen v. a. der Leserin/dem Leser illustrieren, dass unterschiedliche Parameterwerte zu unterschiedlichen Ergebnissen führen.

Das beste Regressionsmodell für die Daten des ReF (also das beste von allen Regressionsmodellen bezogen auf alle Texträume des ReF) ist dasjenige mit einer Clusterlösung von 13 Clustern, in einem Textraum mit den folgenden Merkmalen: Wahl der Unigramm-Merkmale; Einbeziehung aller Unigramm-Formen im Korpus (Minimalfrequenz von 1), aber Ausschluss von zu häufigen und daher uninformativen Unigramm-Formen (mittels einer Stoppwortliste von 50 Elementen). Dieses Vektorraummodell wurde zudem mittels SVD reduziert. Das Modell erreicht einen MSE von

0,072; das entspricht im Durchschnitt etwa einer Abweichung von 7 % der *wh*-Relativpronomen vom tatsächlich beobachteten Prozentsatz der *wh*-Relativpronomen pro Text. Zum Vergleich: Wenn wir das schlechteste Regressionsmodell für den schlechtesten Textraum gewählt hätten, dann hätten wir einen MSE von 0,129 erhalten: Dies entspricht einer Abweichung von 13 % der *wh*-Relativpronomen pro Text und ist damit doppelt so hoch wie die Abweichung des besten Modells. Nicht zuletzt ist es aufschlussreich, die Leistung des besten Regressionsmodells mit den Clustern aus den ReF-Daten mit einem Regressionsmodell zu vergleichen, das die korpuseigenen Textsortenkategorien als Prädiktoren hat: Ein solches Modell weist einen MSE-Wert von 0,121 auf – also einen Wert, der besser ist als der des schlechtesten Regressionsmodells, aber auch schlechter ist als der Wert des besten Modells. In den Abbildungen 5 und 6 wird der MSE-Wert, der mit den korpuseigenen Textsortenklassifikationen ermittelt wurde, durch die gestrichelte Linie namens „text type“ dargestellt.

Das beste Regressionsmodell für die Daten des GerManC wählt eine optimale Clustering-Lösung mit 7 Clustern. Das beste Vektorraummodell zeichnet sich durch einen Grenzwert der Minimalfrequenz von 10 aus, eine Kombination aus Unigramm- und Bigramm-Merkmalen, die Verwendung einer Stoppwortliste mit 100 Elementen und die Reduzierung des Raums mittels SVD. Dieses Modell erreicht einen MSE-Wert von 0,044, verglichen mit einem MSE-Wert von 0,055 eines Regressionsmodells mit den vordefinierten Textsorten als Prädiktoren. Der Unterschied zwischen den beiden Ansätzen ist nicht so groß wie bei den ReF-Daten, was bedeutet, dass sich die gefundenen Cluster sehr wahrscheinlich mit der vordefinierten Texttypenklassifikation überschneiden.

Zusammenfassend lässt sich sagen, dass die Arbeit mit bottom-up-vektorbasierten Clustern in Bezug auf die Vorhersagegenauigkeit einen objektiven Vorteil gegenüber der Arbeit mit den vordefinierten Textsorten bietet. Wir werden daher die Analyse unserer beiden Korpora mit unseren leistungsfähigsten Regressionsmodellen fortsetzen, mit 13 Clustern für die ReF-Texte und 7 Clustern für die GerManC-Texte.

3 Zur textsortenbedingten Verwendung des *wh*-Relativpronomen in der frühen Neuzeit

3.1 Beschreibung der ermittelten Cluster

Die Abbildungen 7 und 8 zeigen unsere Texträume: für Abbildung 7 basierend auf einer semantischen Vektorraumanalyse mit Unigrammen in Form von 13 Clustern; für Abbildung 8 basierend auf einer semantischen Vektorraumanalyse mit Uni-

grammen und Bigrammen in Form von 7 Clustern (für Details zum Verfahren, s. 2.3 bis 2.5) Im Folgenden werden wir die Abbildungen beschreiben; die Tabellen 7 und 8 geben zudem die wichtigsten Informationen zu den Abbildungen (z. B. die genaue Zusammensetzung der einzelnen Cluster) in tabellarischer Form wieder.

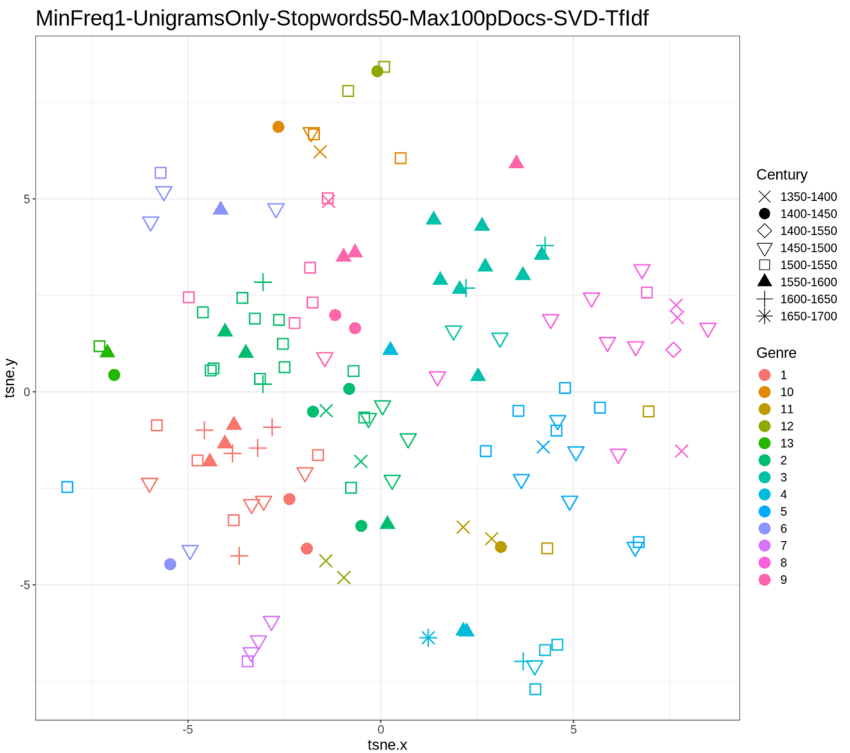


Abb. 7: t-SNE-Diagramm der Ähnlichkeitsmatrix der Texte des ReF (Farbcodierung: Cluster, Formcodierung: Jahrhunderte)

Abbildung 7 zeigt, dass es Unterschiede zur tradtionellen Textsortenklassifikation gibt, die nur 6 Zuordnungen (anstatt 13) vorgenommen hat. Die Textsorte „RE“, die wissenschaftliche Texte beinhaltet, ist in der Clusteranalyse in drei Cluster aufgeteilt: die Cluster 4, 5 und 11. Alle drei Cluster befinden sich in der rechten unteren Hälfte der Abbildung. Die drei Cluster unterscheiden sich in thematischer Hinsicht voneinander: Cluster 4 behandelt eher medizinische Themen (typische Unigramme sind u. a. *pflaster* und *wunde*), Cluster 5 umfasst eher mathematische bzw. rechnerische Unigramme, die bei der Zubreitung von etwas erforderlich sind (z. B. *zahl* und *sieden*) und Cluster 11, das nur 5 Textausschnitte beinhaltet, thematisiert

eher astronomische Themen (typische Unigramme sind *planet* und *stern*). Cluster 4 umfasst einen längeren Zeitraum vom 14. bis ins 17. Jahrhundert, während die beiden anderen Cluster nur bis 1550 gehen, dafür aber ein Jahrhundert früher, nämlich 1350, starten. Das könnte man so deuten, dass die Wissenschaftsprosa zunächst Themen aus Astronomie und der Zubereitung von Heilmitteln behandelt, und sich etwas später der Schwerpunkt dann auf medizinische Themen verlagert. Laut Forschungsliteratur entwickeln sich im 16. Jahrhundert Ansätze einer deutschen Wissenschaftssprache (vgl. Polenz 2000: 144), die im 17. und 18. Jahrhundert schließlich zu neuen Textmustern wie Fachsprachen führen (vgl. Polenz 2013: 18). Auch die lateinischsprachige parallele Texttradition mag eine Rolle bei der vergleichsweise späten Herausbildung einer deutschen Wissenschaftssprache gespielt haben (vgl. Klüsener/Grzega 2012: Sp. 1489–1490); in anderen Bereichen wie dem Protestantismus geschah dies hingegen deutlich früher (bereits im 16. Jahrhundert mit Luthers Schriften, vgl. Beutel 2010: 249–251).

Die Rechts- und Geschäftstexte sind ebenfalls in 3 Cluster unterteilt: Cluster 9, 10 und 12 bestehen mehrheitlich aus diesen Textgruppen. Die drei Cluster befinden sich vorwiegend in der oberen Hälfte des Raums. Alle drei Cluster umfassen den Zeitraum von 1350–1550, sodass man davon ausgehen kann, dass die Unterteilung in Cluster nicht aufgrund verschiedener Jahrhunderte erfolgte. Cluster 9 umfasst zwar mehrheitlich Texte mit rechtlichem und geschäftlichen Charakter, allerdings finden sich darin auch kirchlich-theologische Texte, die ebenfalls Fragen der Obrigkeit und des Gerichts thematisieren. Cluster 10 ist weniger im Raum verstreut als Cluster 9, was möglicherweise damit zusammenhängt, dass alle Texte aus dem Rechtsbereich stammen und interessanterweise nur aus dem ostdeutschen Sprachgebiet (inklusive des Böhmischen). Der Begriff *Ader* ist hier auffallend, lässt sich aber damit erklären, dass im Bergbau (aus dem ein Text stammt) der Begriff verwendet wird und dass der Begriff früher auch in der Bedeutung von „Veranlagung, Neigung“ verwendet wurde. Möglicherweise wurde dieser Begriff v. a. im ostmitteldeutschen Sprachraum verwendet, denn in Cluster 12 – das nur aus Texten aus dem ostmitteldeutschen Sprachraum besteht – findet sich der Begriff ebenfalls. Die Aufteilung der rechtlichen Texttraditionen in die eben genannten Untergruppen mag bedingt sein durch Unterschiede in der Lexik (Cluster 9 aus dem Westoberdeutschen, Cluster 10 und 12 aus dem Ostdeutschen). Cluster 10 und 12 liegen im Vektorraum ebenfalls eng beeinander (wobei es bei Cluster 12 noch eine Abspaltung in die untere Hälfte gibt). Möglicherweise wurden hier zwei Cluster gebildet, da der zeitliche Schwerpunkt bei Cluster 10 in der Mitte des 14. bis Mitte des 15. Jahrhunderts liegt, bei Cluster 12 hingegen in der Mitte des 15. bis Mitte des 16. Jahrhunderts.

Auch die Texte mit vorwiegend erbaulichem Charakter wurden in drei Cluster unterteilt: Dazu gehören Cluster 3, 8 und 7. Cluster 3 und 8 finden sich auf der rechten Seite in der oberen Hälfte; gemeinsam ist beiden das Lexem *heilig*. Cluster

7 hingegen ist ein sehr kleines Cluster (4 Texte) und befindet sich links in der unteren Hälfte. Diese drei Cluster lassen sich weniger eindeutig einer bestimmten Texttradition zuordnen, vielmehr sind v. a. Cluster 3 und 8 „Mischungen" (Cluster 7 ist eindeutig mehrheitlich erbaulich, umfasst aber nur 4 Texte). Cluster 8 umfasst vorwiegend Texte mit erbaulichem Schwerpunkt und berichtendem Charakter, während Cluster 3 eher erbauliche Texte mit einem unterhaltenden Schwerpunkt umfasst: Das Lexem *könig* findet sich so auch in Cluster 1, das fast ausschließlich Texte mit unterhaltendem Charakter umfasst. Auch Lexeme wie *ritter* in Cluster 3 und wie *sakrament* und *teufel* in Cluster 8 zeigen auf, dass erbauliche Texte eine unterschiedliche Schwerpunktsetzung haben können: In Cluster 3 dienen die erbaulichen Texte auch der Unterhaltung, in Cluster 8 eher der Information oder der Belehrung. Cluster 7 umfasst nur Passionstexte (sowohl in Handschrift als auch im Druck), womit bestätigt wird, dass unser Modell semantische Gruppierungen bildet. Der Computer kann hingegen nicht deuten, wenn einzelne Texte (wohl aufgrund von falschen automatischen Annotationen) aufgrund von einzelner „f" und „g" zu einem Cluster zusammengefügt werden. Dies liegt bei Cluster 13 vor, das sich aus unterschiedlichen Zeiträumen, Textsorten und Sprachlandschaften zusammensetzt und wo einzig die eben genannten Lexem gemeinsame Grundlage (und wohl Auslöser für diese Clusterbildung) sind. Das Cluster haben wir aus diesem Grund in der Tabelle weggelassen. Cluster 6 findet sich in der oberen Hälfte auf der linken Seite; es ist ein eher kleines Cluster, mit über 70 % berichtenden Texten. Es geht um wichtige Ereignisse im Umfeld von *landgraf, herzog, bischof* und *papst.* Cluster 1 schließlich dominiert die untere Hälfte auf der linken Seite. Es ist auch das größte Cluster mit 18 Texten und interessanterweise lässt sich dieses Cluster auch mit 67 % recht eindeutig dem unterhaltenden Charakter zuordnen. Diese Texte starten im 15. Jahrhundert, der Schwerpunkt liegt auf dem 15. und 16. Jahrhundert. Typische Lexeme sind *könig* und *ritter* sowie *sagen.*

Texte mit einem kirchlich-religiösen Charakter bilden ein eigenes Cluster (Cluster 2), das sich über den gesamten Zeitraum von 1350 bis 1650 verteilt. Ca. ein Drittel der Texte aus diesem Cluster umfassen religiöse Texte mit erbaulichem Schwerpunkt; dies zeigt sich auch darin, dass das Lexem *heilig* sowohl in Cluster 2 als auch in Cluster 3 (Texte mit erbaulichem Schwerpunkt) vertreten ist. Weitere typische Unigramme für Cluster 2 umfassen *kreuz, seele, gnade* und *christus* – also Themenfelder, mit der sich die kirchlich-theologische Literatur (Schwerpunkt Dogmatik) befasst. Diese theoretischen Texte liegen thematisch näher an der erbaulichen Literatur als an Wissenschaftstexten (im Cluster nur mit 8 % vertreten), die sich mit „realen" Themen wie Anleitungen, Kochbücher, Rezeptbücher, Arzneibücher, Naturlehren oder auch Astronomiebücher befasst.

Tab. 7: Eigenschaften der ermittelten Cluster im ReF

Cluster	Bezeichnung Texttradition	Typische Unigramme	Zeitraum Jhd.	Merkmale
1	Texte mit unterhaltendem Charakter	*könig, ritter, sohn, sagen, gehen, salvator*	1400–1650	Gesamt 12 Texte. 67 % der Texte unterhaltender Charakter, 17 % mit berichtendem Charakter, 11 % mit erbaulichem Charakter, der Rest wissenschaftlich/Realientexte. 67 % der Texte aus 1500–1650.
2	Texte mit kirchlich-theologischem Charakter	*heilig, kreuz, seele, christus, gnade*	1350–1650	Gesamt 26 Texte. 58 % der Texte mit kirchlich-theologischem Charakter, 31 % erbaulicher Charakter, 8 % Wissenschafts-/Realientexte, 3 % Rechtstexte. Texte mit kirchlich-theologischem umfassen den gesamten Zeitraum.
3	Texte mit erbaulichem Schwerpunkt und berichtend-christlichem Charakter	*sakrament, anzeigen, kirche, vergeben, erscheinung, teufel, heilig*	1450–1650	Gesamt 12 Texte. 42 % der Texte erbaulicher Charakter, 25 % berichtender Charakter, 17 % mit wissenschaftlichem Charakter, der Rest unterhaltend und kirchlich-theologisch. Schwerpunkt (2/3) im Zeitraum 1550–1600. Alle Texte aus dem Westdeutschen (2/3 aus dem Westmitteldeutschen)
4	Anleitend-informativ: medizinische Themen	*pflaster, wunde, salben, heilen, krankheit*	1450–1700	Gesamt 9 Texte. 100 % davon zu RE gehörend, die meisten Texte (6) aus dem Zeitraum 1500 bis 1600. 5 Texte aus dem Westoberdeutschen, die restlichen verteilen sich auf das Ostober- und Westmitteldeutsche. Thematisch handelt es sich um Wissenschaftsprosa mit medizinischem Schwerpunkt.
5	Anleitend-informativ: (mathematische) Angaben zur Zubereitung	*zahl, zwei, multiplizieren, geben, sieden, mahlen*	1350–1550	Gesamt 13 Texte. 54 % anleitend-informativ (RE), je 23 % Texte mit berichtendem Charakter (CB) und mit offiziellem Charakter (RG). Zeitlicher Schwerpunkt 1450–1550 (12 Texte). Alle Regionen vertreten, jedoch Schwerpunkt auf dem Westdeutschen (9 Texte, davon 5 Westmitteldeutsch). Thematisch handelt es sich um Wissenschaftsprosa mit technischen Angaben zur Zubereitung.
6	Texte mit berichtendem Charakter	*landgraf, herzog, bischof, papst, könig, kaiser, recht*	1400–1600	Gesamt 7 Texte. 71 % der Texte mit chronikalisch/berichtendem Charakter, die restlichen beiden Texte je einmal erbaulich und einmal rechtlicher/geschäftlicher Charakter. Mehrheit der Texte 1450–1500.

Tab. 7 (continued)

Cluster	Bezeichnung Texttradition	Typische Unigramme	Zeitraum Jhd.	Merkmale
7	Texte mit erbaulichem Schwerpunkt und Schwerpunkt Passion	*jesus, salvator, maria, petrus*	1450–1550	Gesamt 4 Texte. 75 % der Texte erbaulicher Charakter, 25 % unterhaltend. Nur Texte aus dem Oberdeutschen, vorwiegend (75 %) Zeitraum 1450–1500.
8	Texte mit erbaulichem Schwerpunkt und unterhaltendem Charakter	*könig, ritter, beichten, heilig, sünde, jesus*	1350–1550	Gesamt 13 Texte. 31 % der Texte erbaulicher Charakter und 31 % der Texte unterhaltender Charakter. 62 % der Texte aus 1450–1550. Alle Texte aus dem Westdeutschen, auch hier die Mehrheit der Texte aus dem Westmitteldeutschen.
9	Texte mit rechtlichem/geschäftlichen Charakter, Schwerpunkt Obrigkeit	*recht, staat, bürger, stadt, ordnung, gericht, obrigkeit*	1350–1600	Gesamt 12 Texte. 58 % der Texte Rechts- und Geschäftstexte (RG), 25 % kirchlich-theologische Texte. Texte mit Rechtscharakter über gesamten Zeitraum vertreten, am häufigsten 1500–1600. Die Hälfte der Texte aus dem Westoberdeutschen.
10	Texte mit rechtlichem/geschäftlichen Charakter	*bürger, rat, stadt, saal, stift, ader*	1350–1550	Gesamt 5 Texte. 100 % der Texte Rechts- und Geschäftstexte (RG). Nur Texte aus ostdeutschem Sprachgebiet, 60 % der Texte aus Zeitraum 1450–1550.
11	Anleitend-informativ: astronomisch/astrologisch	*zeichen, planet, stern, zeit*	1350–1550	Kleines Cluster mit 5 Texten. 80 % der Texte anleitend-informativ (RE), der Rest kirchlich-theologisch. Schwerpunkt auf dem Westdeutschen (80 % der Texte), Texte gleichmäßig über den ganzen Zeitraum verteilt bis 1550. Thematisch handelt es sich um Wissenschaftsprosa mit astrologisch/astronomischem Schwerpunkt.
12	Texte mit rechtlichem/geschäftlichen Charakter,	*ader, bezahlen, rat, stadt, saal*	1350–1550	Gesamt 5 Texte. 60 % der Texte Rechts- und Geschäftstexte. Die restlichen 20 % jeweils kirchlich-theologischer und unterhaltender Text. Alle Texte aus dem Ostmitteldeutschen, 60 % der Texte aus Zeitraum 1350–1450.

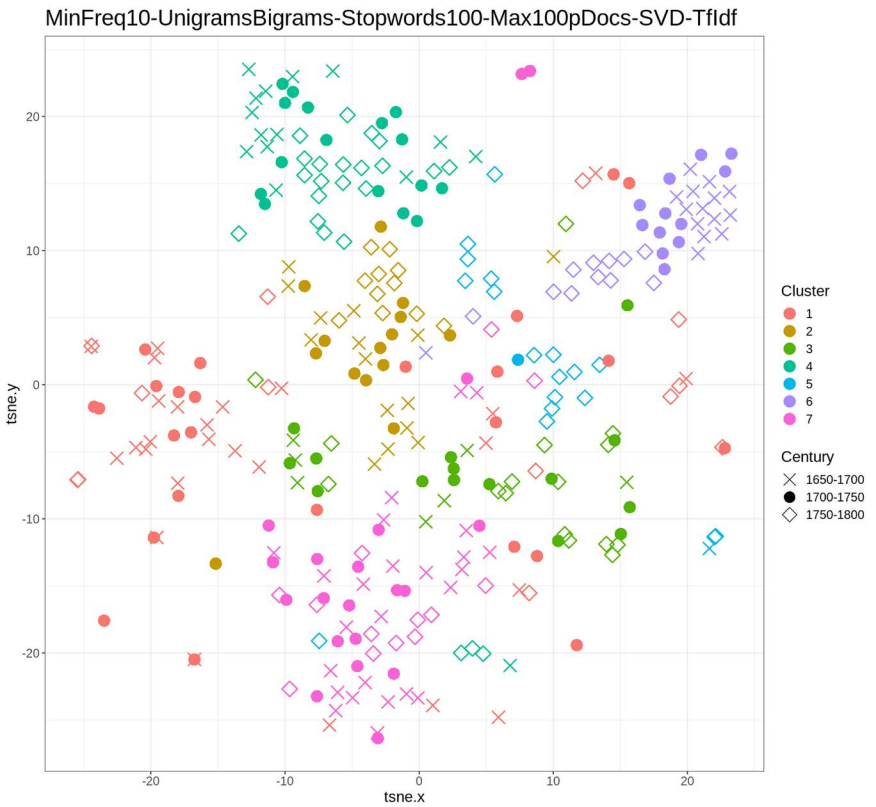


Abb. 8: t-SNE-Diagramm der Ähnlichkeitsmatrix der Texte des GerManC (Farbcodierung: Cluster, Formcodierung: Jahrhunderte)

Abbildung 8 zeigt die Clusteranalyse des GerManC. In dieser Analyse finden sich große Überschneidungen zwischen der Darstellung nach Textsorten und nach Clustern: Alle sieben Cluster haben einen der mittels philologischer Analyse festgestellten Schwerpunkt. Gleichzeitig besteht allerdings keines der Cluster ausschließlich aus Texten aus einer philologisch ermittelten Textsorte. Die Cluster zeigen demnach, so könnte man interpretieren, besser Übergänge und Ähnlichkeiten zwischen einzelnen Textsorten auf. Keines der ermittelten Cluster weist sprachgeografische Besonderheiten auf (obwohl wir mit Wortformen gearbeitet haben und nicht mit Lemmata), denn die Texte der Cluster sind innerhalb der Cluster jeweils recht ausgewogen auf alle fünf Sprachregionen (West- und Ostoberdeutsch, West- und Ostmitteldeutsch, Norddeutsch) verteilt.

Cluster 1 befindet sich vorwiegend in der unteren Hälfte auf der linken Seite und ist moderat im Raum verstreut (v. a. auf der linken Seite der Darstellung). Die

Streuung ist bedingt durch die Darstellung im Vektorraum und durch die „Reduktion“ der Vektorraummodelle auf nur zwei Dimensionen. Unter diesen Texten finden sich sowohl Rechts- als auch Wissenschaftstexte. Interessant ist, dass die Rechtstexte über den ganzen Zeitraum verteilt sind, die Wissenschaftstexte und Texte mit geisteswissenschaftlichem Inhalt hingegen sind nur bis 1750 belegt. Möglicherweise deutet das darauf hin, dass nach 1750 das Textprofil von Rechts- und Geschäftstexte (noch) stärker profiliert ist. In der Lexik sind für die Rechtssprechung und Geschäftssprache typische Unigramme wie *gericht, bezahlen, lohn, sollen* vertreten. Sehr klar abgetrennt und nicht im Raum verstreut ist Cluster 4, das zu knapp 90 % aus unterhaltend-dramatischer Literatur mit dialogischem Charakter besteht. Der dialogische Charakter ist erkennbar an Interjektionen wie *ha* oder *ha ha* sowie *ja* und dem häufigen Auftreten der 2. Person Singular (*bist*), die nicht von der Stoppwortliste herausgefiltert wurde. Thematisch geht es hier um familiäre Beziehungen und Liebesbande. Polenz (2013: 18) spricht von einer belletristischen Literatursprache, die sich im 17. und 18. Jahrhundert entwickelt.

Auch Cluster 2 umfasst unterhaltende Texte, allerdings zeichen sich diese durch einen monologischen Charakter aus. Im Vektorraum grenzt Cluster 2, das zentraler liegt, jedoch an Cluster 4 an, das sich im oberen Drittel befindet. Im Gegensatz zu Cluster 4 finden sich hier flektierte Verbformen in der dritten Person und im Präteritum – was auch bei erzählender Prosa erwarten würde. Thematisch geht es ebenfalls um familiäre Themen, wie Unigramme wie *vater, mutter, fräulein* sowie Bigramme wie *mein bruder, mein vater* zeigen. Knapp 15 % der Texte behandeln geisteswissenschaftliche Themen: Im Vektorraum ist diese Nähe auch dadurch erkennbar, dass Cluster 5 – das kleinste Cluster mit nur 20 Textausschnitten – an die unterhaltende Literatur angrenzt. Dieses Cluster ist gleichzeitig auch dasjenige, das aus den meisten unterschiedlichen Texttypen besteht: Nur 45 % der Texte (aber immer noch der größte Anteil an Texten) umfasst geisteswissenschaftliche bzw. schöngeistige Themen, die die *Natur, Kunst*, *Gefühl* und auch nationale Themen wie das Deutschsein (erkennbar am Unigramm *deutsche*) behandeln. Auch der Rest umfasst unterschiedliche Texte mit berichtendem Charakter sowie aus den Bereichen der Wissenschaft, des Rechts und der Unterhaltung. Das Cluster zeigt, dass die Textsorte *humanities* weniger eindeutig ist, als die philologische Zuordnung vorgibt. Auffällig an diesem Cluster ist, dass 90 % der Texte aus dem Zeitraum 1750–1800 stammen. Dies liese sich so interpretieren, dass diese Textsorte „Geisteswissenschaften“ oder auch Literatur, die sich den (schönen) Künsten widmet, erst in der zweiten Hälfte des 18. Jahrhunderts entstanden ist. Wie Cluster 4 befindet sich Cluster 6 auch gut erkennbar als eigenes Cluster im Vektorraum, und zwar im oberen Drittel auf der rechten Seite. Cluster 6 umfasst fast ausschließlich Texte mit berichtendem Charakter, die in der Semantik v. a. auf den Bericht kriegerischer Handlungen schließen lassen. Dieses Textfeld scheint offenbar für berichtende Texte wie Zeitungen

typisch zu sein. Berichtende Texte des Typus „Zeitung“ wiederum können bereits ab dem 17. Jahrhundert als eigenständige Texttradition ausgemacht werden: Laut Forschungsliteratur erscheinen wöchentlich Zeitungen (bzw. Vorläufer davon) bereits zu Beginn des 16. Jahrhunderts (vgl. Polenz 2000: 140; Polenz 2013: 18). Es ist also plausibel anzunehmen, dass sich im 17. Jahrhundert eine eigenständige Texttradition „Zeitungen“ ausgebildet hat.

Cluster 3 wiederum grenzt an mehrere Cluster an und befindet sich ungefähr in der Mitte des Vektorraums, v. a. an Cluster 1, 2 und 7 angrenzend. Cluster 3 umfasst zu zwei Drittel Wissenschaftsprosa und zu einem Fünftel geisteswissenschaftliche Texte. Die meisten der geisteswissenschaftlichen Texte stammen aus dem Zeitraum 1650–1750 – was bestätigen würde, dass sich diese Textgattung erst spät im 18. Jahrhundert (zumindest inhaltlich durch durch eine charakteristische Lexik) etabliert. Typische Unigramme für diese Diskurstradition sind *wasser, materie* und *feuer*. Cluster 7 ist schließlich, nach Cluster 1, das zweitgrößte Cluster. Im Vektorraum ist es zu großen Teilen klar als eigenes Cluster zu erkennen, es befindet sich in der unteren Hälfte mit einer meist dichten Verteilung. Interessanterweise setzt sich dieses Cluster aber nicht so eindeutig zusammen, sondern ähnelt vielmehr Cluster 5, das sich ebenfalls nicht eindeutig einer mittels philologischen Methoden ermittelten Textsorte zuordnen lässt: Cluster 6 umfasst 36 % mit predigenden-belehrenden Texten und ein Fünftel geisteswissenschaftliche Texte. Anders formuliert scheint es hier eine Texttradition zu geben, die sich mit christlich-theologischen Inhalten, verbunden mit geisteswissenschaftlichen Themen, auseinandersetzt, und zwar mit wissenschaftlicher Prägung, denn immerhin noch 10 % der Texte umfassen Wissenschaftsprosa. Diese Texttradition, so könnte man interpretieren, ist einerseits bedingt durch frühere christlich-theologische Traditionen, gleichzeitig scheint sich eine Öffnung hin zu angrenzenden Diskursen zu ergeben, möglicherweise bedingt durch die Reformation, die die bisherige klerikale Tradition hinterfragt und den Blick vermehrt auf die Leserschaft richtet. In diesem Cluster finden sich vorwiegend religiöse Texte (Predigten) sowie Wissenschaftsprosa aus dem 17. und 18. Jahrhundert. Diesen Befund könnte man so deuten, dass Predigten im 17. und 18. Jahrhundert einen gelehrtensprachlichen Duktus (bzw. Lexik) haben und daher zusammen mit wissenschaftlichen Texten ein Cluster bilden können.

Tab. 8: Eigenschaften der ermittelten Cluster im GerManC

Cluster	Bezeichnung Texttradition	Typische Uni- und Bigramme	Zeitraum Jhd.	Merkmale
1	Texte mit geschäfts-/rechtlichem Charakter	*vermögen, sollen, gericht, bezahlen, gold, erbschaft, lohn*	1650–1800	Gesamt 65 Texte. 62 % der Texte mit rechtlichem Charakter. 15 % mit Wissenschaftsprosa, 12 % mit geisteswissenschaftlichem Inhalt, 5 % mit berichtendem Charakter, je 3 % mit erzählend-unterhaltendem Charakter und mit predigendem Charakter. Texte mit rechtlichem Charakter im gesamten Zeitraum,
2	Texte mit unterhaltendem-monologischen Charakter	*fräulein, mein vater, frauenzimmer, vater, mutter, mein bruder, magd, frau, sprach, rief, sagte*	1650–1800	Gesamt 41 Texte. 85 % der Texte mit unterhaltend-monologischem Charakter, 14 % der Texte mit geisteswissenschaftlichem Charakter. 2 % der Texte mit Wissenschaftsprosa.
3	Texte mit wissenschaftlichem Inhalt	*krankheit, gold, feuchtigkeit, wasser, materie, meer, feuer*	1650–1800	Gesamt 38 Texte. 66 % der Texte mit wissenschaftlichem Charakter, 20 % geisteswissenschaftliche Texte, je 5 % Rechts- und Zeitungstexte, 4 % predigend-belehrender Text.
4	Texte mit unterhaltendem (dialogischen) Charakter	*bist, dir, ha, mann, tochter, sohn, könig, vater, liebe, ha ha, ja*	1650–1800	Gesamt 54 Texte. 87 % der Texte mit unterhaltend dialogischen Charakter, 7 % predigender Charakter, 4 % mit geisteswissenschaftlichem Charakter, 2 % mit berichtendem Charakter.
5	Texte mit schöngeistigem Inhalt	*sprache, gefühl, natur, kunst, deutsche*	v. a. 1750–1800	Gesamt 20 Texte. 45 % der Texte mit geisteswissenschaftlichem Thema, 20 % mit berichtendem Charakter, je 10 % mit Wissenschaftsprosa, Rechtsprosa und erzählende Literatur, 5 % mit predigendem-belehrenden Charakter. 90 % der Texte aus dem Zeitraum 1750–1800.

Tab. 8 (continued)

Cluster	Bezeichnung Texttradition	Typische Uni- und Bigramme	Zeitraum Jhd.	Merkmale
6	Texte mit berichtendem Charakter	*majestät, könig, schiffe, gefangen, kaiserliche, feind, england, regiment*	1650–1800	Gesamt 38 Texte. 95 % der Texte berichtenden Charakter, je 3 % mit Wissenschaft und Geisteswissenschaft. Gleichmäßig über alle Zeiträume verteilt.
7	Texte mit christlich-belehrendem Charakter	*herrn, sünde, gott, dein herz, wahrheit, mutter, Christus, herrlichkeit*	1650–1800	Gesamt 58 Texte. 36 % christlich-predigende Texte, mit 21 % geisteswissenschaftliche Texte, 10 % mit wissenschaftlichen Texten, 3 % unterhaltend-monologisch, 2 % Rechtstext.

3.2 Angaben zur textsortenbedingten Verwendung des *wh*-Relativpronomens aus der Forschungsliteratur

Das *wh*-Relativpronomen wird flektiert und nimmt Bezug auf einen nominalen Ausdruck im übergeordneten Satz; es fungiert als Attribut zu diesem Kopfnomen (vgl. Holler 2013: 266): Im Beispielsatz *Das Kind, **welches** sich über das Geschenk freut* nimmt *welches* Bezug auf die NP *das Kind* im übergeordneten Satz und beschreibt es mithilfe des Relativsatzes näher. Mehrere Untersuchungen belegen die textsortenspezifische Verwendung des *wh*-Relativpronomens in der frühen Neuzeit: Nach ersten, sporadischen Belegen im 15. Jahrhundert gilt es im 16. Jahrhundert als Merkmal der Amts-, Geschäfts- und Gelehrtensprache (vgl. Ebert 1986: 161, Polenz 2013: 302, Reichmann/Wegera 1993: 446). In Dialogen fehlt es im 16. Jahrhundert oft, in Streitschriften und Predigten ist es belegt und am häufigsten tritt es in amtlichen Dokumenten auf (vgl. Ebert 1986: 161). Im 17. Jahrhundert wird das *wh*-Relativprononen als typisch für die gehobene deutschen Bildungssprache charakterisiert (vgl. Ebert 1986: 161, Polenz 2013: 302, Reichmann/Wegera 1993: 446). Das *wh*-Relativpronomen ist sehr häufig in wissenschaftlichen Texten zu finden, vor allem im 17. Jahrhundert, auch tritt es im 16. und 17. Jahrhundert häufig in berichtenden Textsorten und Predigten auf (vgl. Moser 2024). Pickl (2020: 248–251) stellt fest, dass das *wh*-Relativpronomen im Bereich der unterhaltenden Literatur bereits zu Beginn des 17. Jahrhunderts zurückgeht. In Textsorten wie Zeitungen und Wissenschaftsprosa wird es dagegen erst in der 2. Hälfte des 19. Jahrhunderts zurückgedrängt (vgl. Pickl 2020: 248–251). Heute gehört das *wh*-Relativpronomen „vornehmlich der geschriebenen Standardsprache an“ (Dudengrammatik 2022: 752, Paragraf 1314). Neben dem *wh*-Relativpronomen, das im Deutschen erstmals ab dem 15. Jahrhundert belegt ist (vgl. Moser 2023: 471), gibt es noch das *d*-Relativpronomen: Untersuchungen zeigen, dass dieses Pronomen der *default*-Marker bei Relativsätzen ist, d. h., das *d*-Pronomen ist sehr viel häufiger belegt als das *wh*-Relativpronomen (vgl. Moser 2023: 471; Moser 2024: 119). Der Relativsatz kann darüber hinaus auch durch eine Relativpartikel wie *so* oder ein Pronominaladverb wie *dabei* eingeleitet werden; allerdings ist diese Form der Relativsatzeinleitung im frühen Neuhochdeutsch seltener als das *wh*-Relativpronomen belegt (vgl. Moser 2024: 119).

Während die oben erwähnten Untersuchungen zum *wh*-Relativpronomen auf der Arbeit mit vordefinierten Textsorten beruhen, schauen wir uns nun an, welche Vorhersagen zum registerbedingten Auftreten des *wh*-Relativpronomens unsere Regressionsanalyse trifft.[13] Basierend auf den Angaben aus der Forschungsliteratur

13 Weitere Faktoren, die das Auftreten eines *wh*-Relativpronomens beeinflussen können und die in der Analyse nicht berücksichtigt werden, können z. B. sein: Schreiblandschaft, Vermeidung

erwarten wir, dass das *wh*-Relativpronomen ab dem 15. Jahrhundert zunimmt und seinen Gebrauchskontext von amtlichen Kontexten ausweitet auf weitere Textsorten bzw. Textdomänen. Ab dem 17. Jahrhundert ist eine Abnahme des *wh*-Relativpronomens zu erwarten, wobei es Unterschiede zwischen den Textsorten gibt: Bei Texten aus unterhaltenden Traditionen erwarten wir eine frühere Abnahme als in Wissenschaftsprosa oder auch Zeitungen. Bezogen auf unsere Korpora und die Clusteranalyse bedeutet das: Wir erwarten im ReF vorwiegend Zunahmen in den einzelnen Clustern, während im GerManC mit einem Rückgang, mit Unterschieden zwischen den einzelnen Clustern, zu rechnen ist.

3.3 Interpretation der Regressionsanalyse

Die Abbildungen 9 und 10 treffen Vorhersagen darüber, in welchen Clustern das *wh*-Relativpronomen im Zeitraum von 1350 bis 1800 wie häufig auftritt. Die Angaben in Prozent geben die Häufigkeit des *wh*-Relativpronomens (Lemma) unter allen Relativsatzmarkern an. Abbildung 9 zeigt uns die Entwicklung basierend auf den Daten aus dem ReF: Wir können erkennen, dass es in fast allen Clustern, wie zu erwarten, eine Zunahme zu beobachten ist. Ausnahmen bilden Cluster 11 und Cluster 5 mit anleitend-informativem Charakter, die allerdings beide bereits 1550 enden. In beiden Clustern erfolgt die Abnahme auf einem bereits hohen Wert von ca. 70 %. Cluster 4 hingegen, das wie Cluster 5 und 11 Texte mit wissenschaftssprachlichem Charakter hat (jedoch mit inhaltlichem Schwerpunkt auf medizinischen Themen), weist weder eine Zu- noch Abnahme auf, sondern verharrt gleichbleibend auf hohem Niveau. Das könnte man so interpretieren, dass die Wissenschaftsprosa von Anfang an eine vergleichsweise hohe Anzahl an *wh*-Relativpronomen besitzt, es jedoch Unterschiede innerhalb dieser Textsorte gibt, abhängig vom thematischen Schwerpunkt. In den Texten mit rechtlich-geschäftlichem Charakter (Cluster 9, 10, 12) finden wir ebenfalls eine Zunahme (Cluster 10 und 12) oder ein Verharren auf konstant hohem Niveau (Cluster 9). Auffällig ist Cluster 1, das unterhaltende Literatur umfasst: Hier sehen wir von Anfang an ein hohes Niveau (über 70 %), so wie auch bei manchen Rechtstexten (z. B. Cluster 12), wo wir es aber eher erwarten würden. Dieser Befund hinterfragt zumindest teilweise die Annahme, wonach das *wh*-Relativpronomen am Anfang auf offizielle Texte wie Amts- und Geschäftstexte beschränkt ist: Dies jedoch nur teilweise, als dass Cluster 8 – das auch unterhaltende Texte umfasst, jedoch mit erbaulichem Schwerpunkt – sich im erwartet niedrigeren

---

von Gleichklang mit anlautendem Artikel, Präferenzen einzelner Schreiber, Verbstellung, vorhergehende Relativpronomen (Vermeidung von Wiederholungen).

Bereich (meist 20–30 %) bewegt. Ein Vergleich mit den Zahlen aus den „originalen" Textsorten (ebenfalls wieder die Häufigkeit des *wh*-Relativpronomens (Lemma) unter allen Relativsatzmarkern) ist hier kaum aussagekräftig, da die klassische Textsortengliederung mittels der Clusteranalyse weiter ausdifferenziert wurde. Der Vollständigkeit halber haben wir dennoch eine Tabelle angefertigt, sie jedoch in die Fußnote gesetzt:[14] Die Tabelle zeigt die Prozentzahlen für die originalen Textsorten; in der Spalte „Cluster" werden die in etwa dazu entsprechenden Cluster genannt. Wie erwartet zeigt der Vergleich, dass es Unterschiede zwischen den vorhergesagten Frequenzen laut Regressionsanalyse und den belegten Frequenzen laut klassischer Textsortengliederung gibt. Wir werden sehen, dass ein Vergleich in dieser Hinsicht beim GerManC mehr Sinn macht, da hier große Ähnlichkeiten zwischen Clustering und originaler Textsortenbildungen vorliegen.

Abbildung 10 weist unsere 7 ermittelten Cluster im GerManC auf: Nur in Rechtstexten ist hier auch noch im 18. Jahrhundert eine Zunahme belegt. In allen anderen Clustern ist eine Abnahme zu verzeichnen: In Cluster 2 ist der Rückgang etwas später als laut Forschungsliteratur vorhergesagt (dort bereits im 17. Jahrhundert) zu beobachten. In Cluster 4 finden sich, wie zu erwarten, in der unterhaltend-dialogischen Literatur am wenigsten Belege für *wh*-Relativpronomen: Die vorhergesagten Frequenzen bewegen sich allesamt unter der 40 %-Marke. Der Forschungsliteratur zufolge ist in der Wissenschaftsprosa und den Zeitungen erst im 20. Jahrhundert ein Rückgang zu beobachten: Im Gegensatz dazu zeigt unsere Analyse bereits einen Rückgang im 18. Jahrhundert: In Cluster 6 (berichtende Texte) ist dies sehr deutlich zu sehen, aber auch in Cluster 3 (Wissenschaftsprosa) ist ein Rückgang ab 1750 zu erkennen.

14

	1350–1400	1400–1450	1450–1500	1500–1550	1550–1600	1600–1650	Cluster
RG	5 %	6 %	17 %	76 %	47 %	k. Quelle	9,10,12
CB	k.Quelle	0 %	13 %	12 %	66 %	72 %	6
RE	0 %	0 %	4 %	11 %	38 %	83 %*	4,5,11
UN	0 %	0 %	2 %	100 %	39 %	100 %	1
EB	0 %	3 %	2 %	11 %	38 %	37 %	3,7,8
KT	0 %	0 %	2 %	39 %	67 %	100 %	2

Die Tabelle zeigt die Anzahl der *wh*-Relativpronomen nach der „originalen" Textsortengliederung (ReF).

* Und für den Zeitraum 1650–1700 in dieser Textsorte: 80 von 103= 77,6 %.

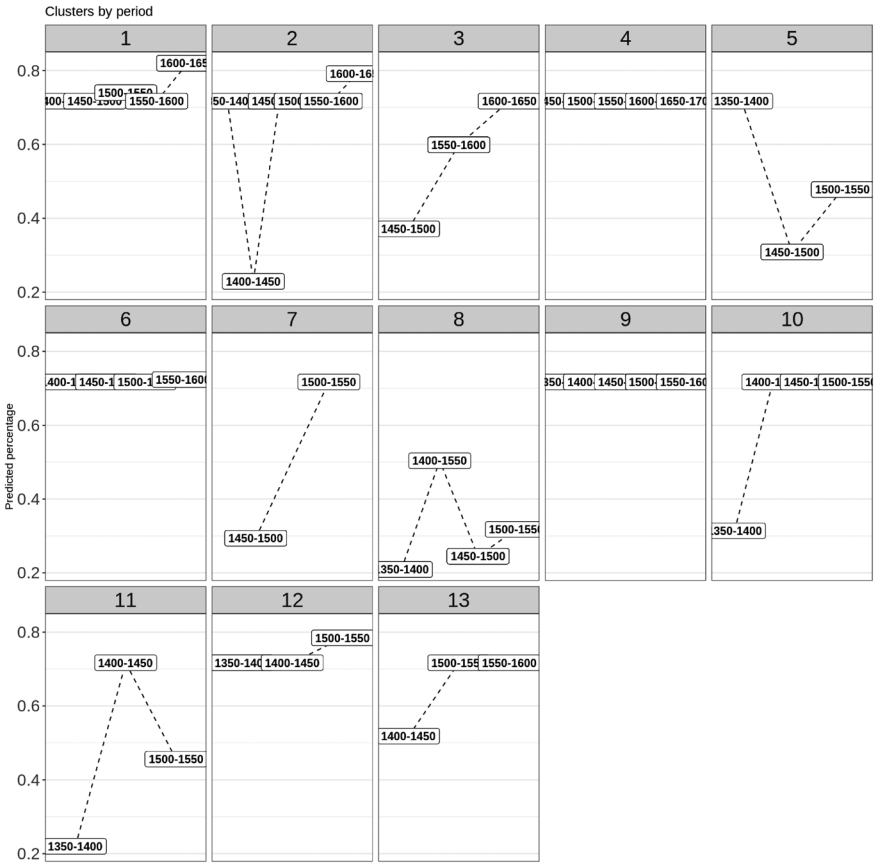


Abb. 9: Vorhergesagte Häufigkeiten des *wh*-Relativpronomen in der frühen Neuzeit mittels der LASSO-Regression (ReF)

Im Gegensatz zum Vergleich der Regressionsanalyse mit originalen Textsorten im ReF verhalten sich die vorhergesagten Frequenzen laut Regressionsanalyse und die belegten Frequenzen laut klassischer Textsortengliederung im GerManC ziemlich ähnlich (vgl. Tabelle 9).[15] Dies war auch zu erwarten, da die Cluster in sehr großen Teilen mit der klassischen Textsortenzuordnung übereinstimmen. Die Tendenzen in Bezug auf Zu- und Abnahme des *wh*-Pronomens stimmen überein, auch wenn die

15 Tabelle 9 zeigt, wie jene in Fußnote 14, die Frequenzen des *wh*-Relativpronomens unter allen Relativsatzmarkern an; die Spalte gibt die entsprechende Zahl des Clusters an, das am ehesten mit der jeweiligen Textsorte übereinstimmt.

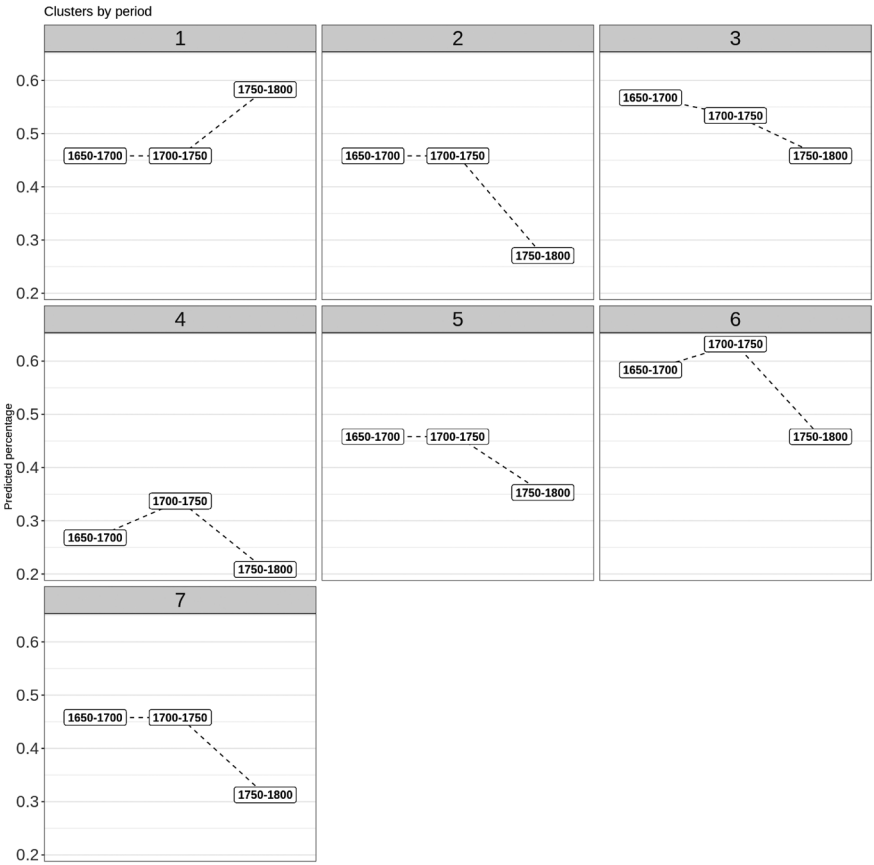


Abb. 10: Vorhergesagte Häufigkeiten des *wh*-Relativpronomen in der frühen Neuzeit mittels der LASSO-Regression (GerManC)

exakten Werte sich um bis zu 15 % unterscheiden können: Cluster 2 startet beispielsweise im ersten Zeitraum bei ca. 45 %, während die Textsorte *narrative* hier bei 60 % beginnt. Im zweiten Zeitraum finden sich im Cluster immer noch 45 %, während in *narrative* hingegen der Wert auf 38 % abgesunken ist. In Cluster 6 wiederum, Textgruppen mit berichtendem Charakter, liegt der erste Zeitraum auf knapp 60 %, während *newspaper* 49 % aufweist. Im zweiten Zeitraum erfolgt jedoch bei beiden Herangehensweisen eine leichte Zunahme von ca. 5 %, worauf wiederum eine Abnahme erfolgt: Das Cluster zeigt im dritten Zeitraum ca. 45 %, die klassische Textzuordnung liegt bei 36 %.

Tab. 9: Anzahl der *wh*-Relativpronomen nach der „originalen“ Textsortengliederung (GerManC)

	1650–1700	1700–1750	1750–1800	Cluster
Legal	28 %	31 %	49 %	1
Narrative	60 %	38 %	16 %	2
Science	36 %	45 %	41 %	3
Drama	10 %	18 %	7 %	4
Humanities	29 %	43 %	24 %	5
Newspaper	49 %	54 %	36 %	6
Sermon	38 %	33 %	21 %	7

4 Zusammenfassung

In diesem Beitrag haben wir einen quantitativen Blick auf die üblicherweise mit philologischen Methoden arbeitenden Textsortenklassifizierung geworfen: Mithilfe von Vektorraummodellen, Clusterbildung und einer Regressionsanalyse haben wir Texte aus zwei Korpora der frühen Neuzeit – dem ReF und dem GerManC – in Form von inhaltlich charakterisierte Textgruppen neu geordnet: VRM können Texte auf Basis von semantischen Ähnlichkeiten in einem dimensionalen Raum anordnen, wobei diese Anordnung auf einem kontinuierlichen Maß (Abstandsindex) erfolgt. Klassische Textklassifikationen können dagegen die Tatsache, dass es mehr oder weniger prototypische Vertreter einer Textsorte gibt und dass es Überschneidungen und Divergenzen in der Textsortenentwicklung gibt, nicht oder nur bedingt widerspiegeln. Um die optimale Anzahl von Clustern zu bestimmen, wurden die Cluster als Prädiktoren in einer Regressionsanalyse genutzt, die als abhängige Variable die Frequenz des *wh*-Relativpronomens vorhersagt (das *wh*-Relativpronomen wird bekanntermaßen textsortenabhängig verwendet). Unsere Analyse zeigt zum einen, dass ein quantitativer Zugang philologische Analysen bzgl. der Textsortenklassifikation bestätigt (dies ist v. a. am GerManC ersichtlich). Gleichseitig liefert unsere Analyse aber auch wichtige Hinweise dahingehend, als dass gerade im 14. bis 16. Jahrhundert Textsorten möglicherweise weniger starr und einheitlich sind, als die klassischen Textsortenzuordnungen aufzeigen: Unsere Analyse weist hier mehr Cluster auf, als die Textsortenklassifikation vorgibt.

Ein „Nachteil“ unseres hier verwendeten quantitativen Zugangs mittels vorwiegend Unigrammen ist, dass eher inhaltlich charakterisierte Textgruppen als durch feste Formeln und syntaktische Muster charakterisierte Textsorten geclus-

tert werden – insofern steht der Vergleich der beiden Herangehensweisen unter einem gewissen Vorbehalt. Vor diesem Hintergrund sollte unsere Studie als Ergänzung (und nicht als Ersatz) der klassischen Textsortengliederung gesehen werden; zudem kann sie als Ausgangspunkt für weitere Studien dienen, die Textsorten quantitativ erforschen möchten (z. B. mittels Ngrammen und mithilfe von geeigneteren Korpora). Dies erscheint unserer Meinung nach besonders lohnenswert für die frühe Neuzeit, die eben dadurch charakterisiert ist, dass neue Textsorten entstehen bzw. bestehende sich weiterentwickeln.

Danksagung: Wir danken den beiden anonymen Gutachterinnen und Gutachtern für hilfreiche Kommentare und Verbesserungsvorschläge. Stefano De Pascale wurde teilweise durch das Forschungsprogramm „Change is Key!“ unterstützt, das vom Riksbankens Jubileumsfond (Referenznummer M21-0021) finanziert wird, und durch ein Juniorpostdoktorandenstipendium von der Forschungsstiftung Flandern (FWO; 1281222N).

Literatur

Bentzinger, Rudolf (2000): Die Kanzleisprachen. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (Hrsg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihre Erforschung. 2. Teilband. Berlin/New York: De Gruyter, 1665–1673.

Besch, Werner (2003): Die Entstehung und Ausformung der neuhochdeutschen Schriftsprache/Standardsprache. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (Hrsg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihre Erforschung. 2. Teilband. Berlin/New York: De Gruyter, 2252–2296.

Betten, Anne (1987): Grundzüge der Prosasyntax. Stilprägende Entwicklungen vom Althochdeutschen zum Neuhochdeutschen (= Reihe Germanistische Linguistik 82). Tübingen: Niemeyer.

Beutel, Albrecht (2010): Sprache. In: Beutel, Albrecht (Hrsg.): Luther Handbuch (UTB). 2. Auflage. Tübingen: Mohr Siebeck, 249–257.

Biber, Douglas/Conrad, Susan (2009): Register, genre, and style (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Bird, Steven/Loper, Edward/Klein, Ewan (2009): Natural Language Processing with Python. O'Reilly Media Inc.

Boleda, Gemma (2020): Distributional Semantics and Linguistic Theory. In: Annual Review of Linguistics 6/1, 213–234.

Bubenhofer, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse (Sprache und Wissen 4). Berlin/New York: De Gruyter. DOI: 10.1515/9783110215854.

Bubenhofer, Noah/Scheurer, Patricia (2014): Warum man in die Berge geht. Das kommunikative Muster „Begründen“ in alpinistischen Texten. In: Hauser, Stefan/Kleinberger, Ulla/Roth, Kersten S. (Hrsg.): Musterwandel – Sortenwandel. Aktuelle Tendenzen der diachronen Text(sorten)linguistik (Sprache in Kommunikation und Medien 3). Bern/Berlin u. a.: Peter Lang, 239–268.

Bubenhofer, Noah/Spieß, Constanze (2012): Zur grammatischen Oberflächenstruktur von Kommentaren. Eine korpuslinguistische Analyse typischer Sprachgebrauchsmuster im kontrastiven Vergleich. In: Grösslinger, Christian/Held, Gudrun/Stöckl, Hartmut (Hrsg.): Pressetextsorten jenseits der „News“: Medienlinguistische Perspektiven auf journalistische Kreativität (Sprache im Kontext 38). Bern/Berlin u. a.: Peter Lang, 87–105.

Clark, Stephen (2015): Vector Space Models of Lexical Meaning. In: Lappin, Shalom /Fox, Chris (Hrsg.): The Handbook of Contemporary Semantic Theory. Hoboken: John Wiley & Sons, 493–522. DOI:10.1002/9781118882139.ch16.

Duden-Grammatik (2022): Duden. Die Grammatik. Hrsg. von Angelika Wöllstein und der Dudenredaktion. 10., völlig neu verfasste Auflage. Berlin: Dudenverlag.

Durrell, Martin/Bennett, Paul/Scheible, Silke/Whitt, Richard J. (2012): The GerManC Corpus. Oxford: Oxford Text Archive. <ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2544> (03.07.2024).

DWB = Deutsches Wörterbuch von Jacob und Wilhelm Grimm. 16 Bde. in 32 Teilbänden. Leipzig 1854–1961. Quellenverzeichnis Leipzig 1971. Digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/21. Online abrufbar unter: https://www.woerterbuchnetz.de/DWB.

Ebert, Robert P. (1986): Historische Syntax des Deutschen II: 1300–1750 (Langs Germanistische Lehrbuchsammlung 6). Bern/Berlin u. a.: Peter Lang.

Ebert, Robert P./Reichmann. Oskar/Solms, Hans-Joachim Solms/Wegera, Klaus-Peter (1993): Frühneuhochdeutsche Grammatik (= Sammlung kurzer Grammatiken germanischer Dialekte. A: Hauptreihe. 12). Tübingen: Niemeyer.

Elspaß, Stephan (2008): Vom Mittelneuhochdeutschen (bis ca. 1950) zum Gegenwartsdeutsch. In: Zeitschrift für Dialektologie und Linguistik 75, 1–20.

Feilke, Helmuth (2003): Textroutine, Textsemantik und sprachliches Wissen. In: Linke, Angelika/Ortner, Hanspeter/Portmann-Tselikas, Paul R. (Hrsg.): Sprache und mehr: Ansichten einer Linguistik der sprachlichen Praxis (Reihe germanistische Linguistik 245). Tübingen: Niemeyer, 209–229.

Firth, John R. (1957): Studies in Linguistic Analysis. Oxford: Blackwell.

Fix, Ulla (2008): Textsorte – Textmuster – Textmustermischung: Konzept und Analysebeispiel. In: Fix, Ulla (Hrsg.): Texte und Textsorten: Sprachliche, kommunikative und kulturelle Phänomene (Sprachwissenschaft 5). Berlin: Frank & Timme, 65–81.

Geeraerts, Dirk (2017): Distributionalism, old and new. Each Venture a New Beginning. Studies in Honor of Laura A. Janda. Bloomington, IN: Slavica Publisher, 29–38.

Geeraerts, Dirk/Speelmann, Dirk/Heylen, Kris/Montes, Mariana/De Pascale, Stefano/Franco, Karlien/Lang, Michael (2024): Lexical Variation and Change. A Distributional Semantic Approach. Oxford: Oxford University Press.

Günthner, Susanne/Knoblauch, Hubert (1994): Forms are the Food of Faith. Gattungen als Muster kommunikativen Handelns. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 46, 693–723.

Haaf, Susanne/Schuster, Britt-Marie (2023) (Hrsg.): Historische Textmuster im Wandel. Neue Wege zu ihrer Erschließung (Reihe Germanistische Linguistik 331). Berlin/Boston: De Gruyter.

Hartmann, Peter (1971): Texte als linguistisches Objekt. In: Stempel, Wolf-Dieter (Hrsg.): Beiträge zur Textlinguistik. München: Fink, 9–30.

Hartweg, Frédéric/Wegera, Klaus-Peter (2005): Frühneuhochdeutsch: Eine Einführung in die deutsche Sprache des Spätmittelalters und der frühen Neuzeit. 2., neu bearbeitete Auflage. Tübingen: Niemeyer.

Haugen, Einar (1966): Dialect, nation, language. In: American Anthropologist 68, 922–935.

Heinemann, Wolfgang/Viehweger, Dieter (1991): Textlinguistik. Eine Einführung (Reihe Germanistische Linguistik Kollegbuch 115). Tübingen: Niemeyer. DOI: 10.1515/9783111376387.

Hausendorf, Heike (2023): Die „Allgemeine Zeitung“ und ihre Texte. In: Haaf, Susanne/Schuster, Britt-Marie: Historische Textmuster im Wandel. Neue Wege zu ihrer Erschließung (Reihe Germanistische Linguistik 331). Berlin/Boston: De Gruyter, 205–252.

Herbers, Birgit/Kösser, Sylwia/Lemke, Ilka/Wenner, Ulrich/Berger, Juliane/Kwekkeboom, Sarah/Thielert, Frauke (2021): Dokumentation zum Referenzkorpus Frühneuhochdeutsch und Referenzkorpus Deutsche Inschriften (= BLA 24). Bochum: Bochumer Linguistische Arbeitsberichte.

Holler, Anke (2013): d- und w-Relativsätze. In: Altmann, Hans/Meibauer, Jörg/Steinbach, Markus (Hrsg): Handbuch Satztypen (De Gruyter Lexikon). Berlin/New York: De Gruyter, 266–300. DOI: 10.1515/9783110224832.266.

James, Gareth/Witten, Daniela/Hastie. Trevor/Tibshirani, Robert (2021): An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics). 2. Auflage. New York, NY: Springer.

Kabatek, Johannes (2015): Wie kann man Diskurstraditionen kategorisieren?. In: López Serena, Araceli/Octavio de Toledo, Álvaro/Winter-Froemel, Esme (Hrsg.): Diskurstraditionelles und Einzelsprachliches im Sprachwandel / Tradicionalidad discursiva e idiomaticidad en los procesos de cambio lingüístico (SkriptOralia). Tübingen: Narr, 51–65.

Kästner, Hannes/Schütz, Eva/Schwitalla, Johannes (2000): Die Textsorten des Frühneuhochdeutschen. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (Hrsg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihre Erforschung. 2. Teilband. Berlin/New York: De Gruyter, 1605–1623.

Klüsener, Bea/Grzega, Joachim (2012): Lemma „Wissenschaftsrhethorik“. In: Ueding, Gert (Hrsg.): Historisches Wörterbuch der Rhetorik. Berlin/Boston: De Gruyter, Sp. 1486–1504.

Koch, Peter (1997): Diskurstraditionen. Zu ihrem sprachtheoretischen Status und ihrer Dynamik. In: Frank, Barbara/Haye, Thomas/Tophinke, Doris (Hrsg.): Gattungen mittelalterlicher Schriftlichkeit (SkriptOralia 99). Tübingen: Narr, 43–79.

Köhler, Hans-Joachim (1981): Fragestellungen und Methoden zur Interpretation frühneuzeitlicher Flugschriften. In: Köhler, Hans-Joachim (Hrsg.): Flugschriften als Massenmedien der Reformationszeit. Stuttgart: Klett-Coda, 1–28.

Kuhn, Hugo (1969): Versuch einer Literaturtypologie des deutschen 14. Jahrhunderts. In: Sonderegger, Stefan/Haas, Alois M./Burger, Harald (Hrsg.): Typologia litterarum. Festschrift für Max Wehrli. Zürich: Verlag, 261–280.

Lasch, Alexander (2023): Am Grab und darüber hinaus: Leichenpredigten und Herrnhutische Lebensbeschreibungen im (kognitionslinguistisch-konstruktionsgrammatischen) Vergleich. In: Haaf, Susanne/Schuster, Britt-Marie (Hrsg.): Historische Textmuster im Wandel: Neue Wege zu ihrer Erschließung (Reihe Germanistische Linguistik 331). Berlin/New York: De Gruyter, 419–438.

Lenci, Alessandro (2018): Distributional Models of Word Meaning. In: Annual Review of Linguistics 4/1, 151–171.

Lenci, Alessandro/ Sahlgren, Magnus (2023): Distributional Semantics (Studies in Natural Language Processing). Cambridge: CUP.

Oesterreicher, Wulf (1997): Zur Fundierung von Diskurstraditionen. In: Frank, Barbara/Haye, Thomas/Tophinke, Doris (Hrsg.): Gattungen mittelalterlicher Schriftlichkeit (SkriptOralia 99). Tübingen: Narr, 19–41.

Maaten van Der, Laurens/Hinton, Geoffrey (2008): Visualizing Data using t-SNE. In: Journal of Machine Learning Research 9/86, 2579–2605.

Mattheier, Klaus J. (2003): German. In: Deumert, Ana/Vandenbussche, Wim (Hrsg.): Germanic Standardizations. Past to Present (Studies in Language and Society 18). Amsterdam/Philadelphia: Benjamins, 245–280.

Mazzola, Giulia, Stefano De Pascale & Malte Rosemeyer (2023): Nuevas herramientas en la lingüística diacrónica: tradiciones discursivas y lingüística computacional. In: Cornillie, Mazzola, Giulia/Thegel, Miriam (Hrsg.): Conceptos y aplicaciones. Frankfurt a. M./Madrid: Vervuert Verlagsgesellschaft, 89–118. DOI: 10.31819/9783968694832-005.

Meier, Jörg (2012): Die Bedeutung der Kanzleien für die Entwicklung der deutschen Sprache. In: Greule, Albrecht/Meier, Jörg/Ziegler, Arne (Hrsg.): Kanzleisprachenforschung. Ein internationales Handbuch. Berlin/Boston: De Gruyter, 3–14.

Moser, Ann-Marie (2023): The ups and downs of relative particles in German diachrony: on loss, grammaticalization, and standardization. In: Journal of Historical Linguistics 13/3, 461–487. DOI: 10.1075/jhl.22026.mos.

Moser, Ann-Marie (2024): Korpusanalytische vs. „klassische“ Methode im Vergleich: Beispiel Relativsätze (1350–1700). In: Germanistische Linguistik 2, 103–146.

Oesterreicher, Wulf (1997): Zur Fundierung von Diskurstraditionen. In: Frank, Barbara/Haye, Thomas/Tophinke, Doris (Hrsg.): Gattungen mittelalterlicher Schriftlichkeit (SkriptOralia 99). Tübingen: Narr, 19–41.

Pickl, Simon (2020): Factors of Selection, Standard Universals, and the Standardisation of German Relativisers. In: Language Policy 19, 235–258. DOI: 10.1007/s10993-019-09530-3.

Pilehvar, Mohammad Taher/Camacho-Collados, Jose (2020): Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. In: Synthesis Lectures on Human Language Technologies 13/4, 1–175.

Pfefferkorn, Oliver (2005): Übung der Gottseligkeit. Die Textsorten Predigt, Andacht und Gebet im deutschen Protestantismus des späten 16. und 17. Jahrhunderts (Deutsche Sprachgeschichte 1). Frankfurt (Main): Lang.

Polenz, Peter von (2000): Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart. Band 1 (Einführung, Grundbegriffe, Deutsch in der frühbürgerlichen Zeit). Berlin/New York: De Gruyter.

Polenz, Peter von (2013): Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart. Band 2 (17. und 18. Jahrhundert). Berlin/New York: De Gruyter.

Referenzkorpus Frühneuhochdeutsch (1350–1650) (2021): Korpus diplomatisch transkribierter und annotierter Texte des Frühneuhochdeutschen. Verfasst unter der Leitung von Ulrike Demske, Stefanie Dipper, Klaus-Peter Wegera und Hans-Joachim Solms. <www.linguistics.rub.de/ref/index.html> (30.06.2024).

Reichmann, Oskar (1996): Autorenintention und Textsorte. In: Frank, Barbara/Haye, Thomas/Tophinke, Doris (Hrsg.): Gattungen mittelalterlicher Schriftlichkeit (SkriptOralia 99). Tübingen: Narr, 119–133.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Reichmann, Oskar/Wegera, Klaus-Peter (1988) (Hrsg.): Frühneuhochdeutsches Lesebuch. Tübingen: Niemeyer.

Reichmann, Oskar/Wegera, Klaus-Peter (Hrsg.) (1993): Frühneuhochdeutsche Grammatik. Bearb. von Robert Peter Ebert, Oskar Reichmann, Hans-Joachim Solms und Klaus-Peter Wegera (Sammlung kurzer Grammatiken germanischer Dialekte, A. Hauptreihe, 12). Tübingen: Niemeyer.

Scharloth, Joachim (2017): Ist die AfD eine populistische Partei? – Eine Analyse am Beispiel des Landesverbandes Rheinland-Pfalz. In: Aptum. Zeitschrift für Sprachkritik und Sprachkultur 13/1, 5–19. DOI: 10.46771/9783967691559_1.

Scharloth, Joachim (2023): Wechselnichtigkeiten: Paraphrasenattribution als Methode der Identifikation von Textmustern auf Ansichtskarten. In: Hausendorf, Heiko/Scharloth, Joachim/Sugisaki, Kyoko/Bubenhofer, Noah (Hrsg.): Ansichten zur Ansichtskarte: Textlinguistik, Korpuspragmatik und Kulturanalyse. Bielefeld: transcript, 41–60. DOI: 10.1515/9783839466346-003.

Schoenke, Eva (2000): Textlinguistik im deutschsprachigen Raum. In: Brinker, Klaus/Antos, Gernd/Heinemann, Wolfgang/Sager, Sven F. (Hrsg.): Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung. Linguistics of Text and Conversation. An International Handbook. Berlin/New York: De Gruyter 123–131.

Schnelle, Gohar (2020): Verbstellungsvarianten als Indikator für Narrativität im Deutschen? Eine explorative Studie zur Definition althochdeutscher Register, in: Pasques, Delphine/Wich-Reif, Claudia (Hrsg.): Textkohärenz und Gesamtsatzstrukturen in der Geschichte der deutschen und französischen Sprache des 8. bis zum 18. Jahrhunderts. Akten zum Internationalen Kongress an der Universität Paris-Sorbonne vom 15. bis 17. November 2018 (Berliner Sprachwissenschaftliche Studien 35). Berlin: Weidler Buchverlag, 11–48

Schuster, Britt-Marie (2019): Sprachgeschichte als Geschichte von Texten. In: Bär, Jochen A./Lobenstein-Reichmann, Anja/Riecke, Jörg (Hrsg.): Handbuch Sprache in der Geschichte (Handbücher Sprachwissen 8). Berlin/Boston: De Gruyter, 219–240. DOI: 10.1515/9783110296112-008.

Schuster, Britt-Marie/Haaf, Susanne: Fünf Thesen zur Untersuchung des Textsortenwandels. In: Haaf, Susanne/Schuster, Britt-Marie: Historische Textmuster im Wandel. Neue Wege zu ihrer Erschließung (Reihe Germanistische Linguistik 331). Berlin/Boston: De Gruyter, 15–40.

Schuster, Britt-Marie/Thielert, Frauke/Haaf, Susanne (2023): Fragen stellen in Pressetextsorten. In: Haaf, Susanne/Schuster, Britt-Marie: Historische Textmuster im Wandel. Neue Wege zu ihrer Erschließung (Reihe Germanistsiche Linguistik 331). Berlin/Boston: De Gruyter, 153–204.

Solms, Hans-Joachim (2000): Soziokulturelle Voraussetzungen und Sprachraum des Frühneuhochdeutschen. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (Hrsg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihre Erforschung. 2. Teilband. Berlin/New York: De Gruyter, 1513–1527.

Sonderegger, Stefan (1979): Grundzüge deutscher Sprachgeschichte. Diachronie des Sprachsystems. Bd. 1: Einführung, Genealogie, Konstanten. Berlin/New York: De Gruyter.

Thielert, Frauke/Georgi, Christopher Georgi (2023): Formelhafte Sprache in Pressetexten. In: Haaf, Susanne/Schuster, Britt-Marie: Historische Textmuster im Wandel. Neue Wege zu ihrer Erschließung (Reihe Germanistische Linguistik 331). Berlin/Boston: De Gruyter, 117–152.

Tschirch, Fritz (1989): Geschichte der deutschen Sprache. Bd. II.: Entwicklung und Wandlungen der deutschen Sprachgestalt vom Hochmittelalter bis zur Gegenwart. 3., erg. und überarbeitete Aufl. von Werner Besch. Berlin/New York: De Gruyter.

Van de Velde, Freek/Pijpops, Dirk (2021): Investigating lexical effects in syntax with regularized regression (lasso). In: Journal of Research Design and Statistics in Linguistics and Communication Science 6/2, 166–199.

Wegera, Klaus-Peter/Solms, Hans-Joachim/Demske, Ulrike/Dipper, Stefanie (2021): Reference Corpus of Early New High German (1350–1650) (1.0.2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5793616

Wiesinger, Peter (2012): Bairisch-österreichisch – Die Wiener Stadtkanzlei und die habsburgischen Kanzleien. In: Greule, Albrecht/Meier, Jörg/Ziegler, Arne (Hrsg.): Kanzleisprachenforschung. Ein internationales Handbuch. Berlin/Boston: De Gruyter, 415–440.