BmE

Separatum aus:

THEMENHEFT 12

*Elisabeth Lienert / Joachim Hamm*
*Albrecht Hausmann / Gabriel Viehhauser (Hrsg.)*

# Digitale Mediävistik

## Perspektiven der Digital Humanities für die Altgermanistik

Publiziert im November 2022.



http://www.erzaehlforschung.de – Kontakt: herausgeber@erzaehlforschung.de
ISSN 2568-9967

*Zitiervorschlag für diesen Beitrag:*
Viehhauser, Gabriel: Digitale Methoden der Textanalyse für die Altgermanistik, in: Lienert, Elisabeth/Hamm, Joachim/Hausmann, Albrecht/Viehhauser, Gabriel (Hrsg.): Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik, Oldenburg 2022 (BmE Themenheft 12), S. 203–246 (online).

*Gabriel Viehhauser*

# Digitale Methoden der Textanalyse für die Altgermanistik

*Abstract.* Der Beitrag gibt einen Einblick in einige Methoden der digitalen Textanalyse (Frequenzanalyse, *Principal Component Analysis*, *Lexical Diversity* und *Topic Modeling*), situiert diese in der Theoriediskussion der Digital Humanities und erprobt ihre Anwendung auf mittelhochdeutsche Literatur anhand eines Minnesangkorpus. Zum einem sollen dabei spezifische Herausforderungen herausgestellt werden, die sich bei der Anwendung digitaler Analysemethoden auf mittelhochdeutsche Texte bieten. Zum anderen plädiert der Beitrag dafür, die Methoden in einem multiperspektivischen Zugang zu nutzen, der das Spektrum von der digitalen Makro- bis zur qualitativen Detailanalyse umfasst.

## 1. Einleitung

Der folgende Beitrag will einen kurzen Einblick in einige Methoden geben, die sich zurzeit in den Digital Humanities im Rahmen der digitalen Textanalyse etabliert haben, und an Beispielen aus der mittelhochdeutschen Literatur überprüfen, ob solche Methoden auch in diesem Bereich zum Einsatz kommen können und welche besonderen Probleme damit verknüpft sind.

Letztlich basieren sämtliche Verfahren der digitalen Textanalyse, so vielgestaltig sie auch sein mögen, auf derselben Grundlage, nämlich auf der Auszählung von Features, die in einem Text auftreten, wobei in den meisten Fällen diese Features schlicht die einzelnen Wörter eines Textes darstellen. Die Methoden bewegen sich mithin notwendigerweise auf der Textober-

fläche. Der Computer kann im Rahmen der digitalen Textanalyse also eigentlich nichts anderes tun, als einfach nur Wörter zählen, aber dieses Wörterzählen kann in unterschiedlich elaborierten Formen erfolgen, die es zum Teil sogar erlauben, über die explizit gegebene Textoberfläche hinaus Schritte in Richtung auf komplexere Textbedeutungen hin zu gehen. Das Spektrum reicht dabei von der Erstellung von Konkordanzen und einfachen deskriptiven Statistiken über Verfahren, die in der Korpuslinguistik und im Information Retrieval gängig sind, bis hin zu komplexeren Modellen der distributionellen Semantik und Machine-Learning-Verfahren.

## 2. Methodologische Vorüberlegungen

Bevor ich ein paar dieser Verfahren vorstelle und auf konkrete Beispiele zu sprechen komme, erscheint es mir notwendig, einige methodologische Überlegungen voranzuschicken und die einschlägige Theoriediskussion der Digital Humanities und insbesondere der Digital Literary Studies zu rekapitulieren, da nur so deutlich werden kann, vor welchem Hintergrund sich solche digitalen Methoden bewegen und welcher Geltungsbereich für sie zu veranschlagen ist.

Computer sind als Rechenmaschinen immer dann besonders performant, wenn es darum geht, große Mengen an Daten auszuwerten und dieselben regelhaften Abläufe wiederholt auszuführen. Für die Untersuchung von Einzelfällen ist die Erstellung eines Algorithmus eigentlich überflüssig und wohl auch nicht nachhaltig, denn eine eigenständige Neuprogrammierung lohnt sich erst dann, wenn dem Algorithmus ein gewisser Grad der Generalisierbarkeit zu eigen ist. Softwarelösungen, die sich nur auf einen einzelnen Fall anwenden lassen und danach unbrauchbar werden, werden nur die wenigsten zur Benutzung eines Computers veranlassen.

Schon daraus ergibt sich, dass man bei der Arbeit mit dem Computer dazu tendiert, von Einzelfällen abzusehen und den Blick auf das große Ganze, die Makroperspektive, zu lenken. Mit dieser Fokussierung läuft die

Anwendung von Computern zunächst einer (zumindest unterstellten) grundsätzlichen Ausrichtung der Geisteswissenschaften auf das Individuelle und das Idiographische zuwider.[1] Auch wenn zu fragen ist, ob das Ansetzen einer solchen binären Klassifikation die individualisierenden Züge der Geisteswissenschaften nicht zu stark überbetont, so kann doch der Einsatz des Computers in den Humanities damit leicht zur Provokation werden.

Eine solche Provokation durch den Blickwechsel auf die Makroperspektive wird etwa durchaus bewusst von Franco Moretti in Kauf genommen, der das für die digitale Literaturwissenschaft wohl bislang wirkmächtigste Konzept entwickelt hat, nämlich den Ansatz des Distant Reading, anfangs sogar noch ohne den Blick auf das Digitale, nämlich im Kontext der Weltliteraturdebatte (Moretti 2000). Moretti geht dabei von dem Befund aus, dass die Gesamtmenge an vorhandenen Texten von einem einzelnen Menschen selbst bei größter Anstrengung schlicht nicht überschaut werden kann, da die dafür nötige Lesezeit die Lebenszeit einer Einzelperson bei weitem übersteigt. Daraus ergibt sich, dass bei jeglicher kulturwissenschaftlicher Analyse mit Selektivität und *bias* zu rechnen ist – und etwa Weltliteratur am Ende dann doch wieder vorwiegend als europäisches Konzept gedacht wird (Moretti 2000, S. 55). Moretti plädiert daher dafür, das genaue Lesen von Einzeltexten zugunsten eines kursorischen, sekundären Distant Reading zu verabschieden: Wenn dem Problem der unüberschaubaren Fülle mit immer mehr Lesen schlicht nicht beizukommen ist, dann, so die Überlegung Morettis, sollte man endlich aufhören zu lesen, und statt dessen nach neuen, alternativen Formen der Auswertung von Texten suchen (Moretti 2000, S. 57).

Distant Reading ist damit an das Versprechen eines ehrlicheren, weniger kolonialistisch und kanonistisch geprägten Zugangs geknüpft, der es nach Moretti nicht zuletzt erlauben soll, literaturgeschichtliche Zusammenhänge in ihrer Vollständigkeit zu erfassen und auf Muster gesellschaftlicher Machtstrukturen abzubilden. Wie in der Forschung bemerkt wurde (Underwood 2017), steht Morettis Konzeption damit in einer literatur-

soziologischen Tradition, auf deren Grundlage ästhetische Unterschiede nivelliert werden bzw. sekundär erscheinen und Texte in Hinblick auf ihre Strukturmuster abstrahiert werden, um so gesellschaftliche Machtmechanismen analysieren zu können: »Forms are the abstract of social relationships: so, formal analysis is in its own modest way an analysis of power« (Moretti 2000, S. 66) - neo-marxistische Strukturanalyse also an Stelle ästhetischer Ausdifferenzierung.

Dass mit dem Blick auf die Makroperspektive die Schärfe im Detail verloren geht, war freilich schon Moretti bewusst. Dieser Verlust muss seiner Ansicht nach jedoch als Erkenntnisbedingung für den Blick aufs ›große Ganze‹ in Kauf genommen werden (Moretti 2000, S. 57f.). Ein solcher Verlust an Detailschärfe ist aber nun eigentlich auch nicht spezifisch für den Einsatz digitaler Methoden, sondern ein erkenntnistheoretischer *trade-off*, der eigentlich immer einzukalkulieren ist, wenn es etwas zu erkennen gilt: Wenn man etwa von weit oben auf die Erde blickt, dann lässt sich gut der Zusammenhang der Kontinente erkennen; die Details der Erdoberfläche werden jedoch vernachlässigt. Fokussiert man hingegen auf die Details, bleiben die größeren Zusammenhänge aus dem Blick.

Letztlich, so ist in der Theoriediskussion der Digital Humanities der letzten Jahre immer wieder betont worden (vgl. in Auswahl: Ciula [u. a.] 2018; McCarty 2004; Flanders/Jannidis 2018; Piper 2017), beruhen Erkenntnisse auf Modellierungen, und Modelle (im Sinne von Stachowiak 1973) sind zwar immer Abbilder von etwas, von Sachverhalten oder der Wirklichkeit, aber nicht die Sache selbst. Sie sind daher notwendigerweise Verkürzungen, lassen bestimmte Details beiseite und heben bestimmte hervor. Modelle werden zudem zu einem bestimmten Zweck zum Einsatz gebracht - und dieser Zweck lässt sich überhaupt nur deswegen erfüllen, gerade weil Details beiseite gelassen und Ansichten akzentuiert werden.

Daraus folgt aber wiederum, dass kein Modell die Wirklichkeit in ihrer Komplexität vollständig abzubilden vermag; im Gegenteil, mit einem prominenten Aphorismus des Statistikers George Box ließe sich sogar umge-

kehrt sagen, dass alle Modelle letztlich falsch und defizitär sind: »all models are wrong«. Dass man sie dennoch zum Einsatz bringen möchte, liegt schließlich wieder an ihrer pragmatischen Komponente: »all models are wrong, but some are useful« (Box 1976, S. 201).

Es ist in den Digital Humanities schon öfter bemerkt worden, dass sich eine solche Vorstellung nun durchaus mit geisteswissenschaftlichen Idealen wie Komplexität und Multiperspektivität vereinbaren lässt (beispielsweise bei So 2017; Piper 2017; Pierazzo 2018; Underwood 2020; Viehhauser 2020): Es gibt eben nicht nur eine ›objektive‹ Sichtweise, sondern im Gegenteil lediglich unterschiedliche Blickpunkte. Gerade die besondere Bedeutung, die Modellierungen in den digitalen Geisteswissenschaften zugemessen wird, hat in besonderem Maße das Bewusstsein für diesen Sachverhalt geschärft. Mit dem Modellbegriff dürften sich wohl auch Erkenntnisprozesse in den ›traditionellen‹ Geisteswissenschaften beschreiben lassen, und es scheint paradox, dass dort das Bewusstsein für die Verkürzungen des eigenen Standpunkts mitunter sogar weit weniger ausgeprägt erscheint als in den technikgeprägten Digital Humanities (vgl. Piper 2016).

Dass der Rekurs auf den Modellierungsgedanken für die digitalen Geisteswissenschaften besonders naheliegend ist, ergibt sich wohl schon aus der Sache selbst: Schon per Definition besteht die Digitalisierung eines analogen Objekts (etwa eines Analogsignals) darin, dass dieses in diskrete Einheiten aufgelöst wird, die den analogen Zustand immer nur annähernd wiedergeben können, dafür aber zähl- und berechenbar sind. Ein Digitalsignal wäre so gesehen ein verkürzendes Modell eines Analogsignals, das zum Zweck der besseren Verarbeitung erstellt wird.

Zudem müssen Modelle in den Digital Humanities – und damit unterscheiden sich diese von den traditionellen Geisteswissenschaften – notwendigerweise immer auch formalisiert sein (Jannidis/Flanders 2018, S. 28; Jannidis 2018, S. 99). Erst durch die genaue Definition der einzelnen Komponenten und Relationen wird es möglich, Modelle zur Weiterverarbeitung an den Computer zu übergeben. Modellen in den digitalen

Geisteswissenschaften eignet daher ein höherer Grad an Explizitheit an, als dies in den traditionellen Geisteswissenschaften der Fall ist.[2]

Dies sollte jedoch keinesfalls zur Annahme verführen, dass es in Modellen der digitalen Geisteswissenschaften keine blinden Flecken gäbe. So tendieren Ansätze in den Digital Humanities oftmals dazu, tieferliegende Fragestellungen nicht direkt, sondern unter Zuhilfenahme von beobachtbaren Indikator- oder instrumentellen Variablen zu bearbeiten. (Moretti 2013, S. 2). Ein Beispiel wäre hier der Schluss von Worthäufigkeiten (auf der Textoberfläche) auf tieferliegende Bedeutungsschichten, etwa vom Auftreten von Begriffen auf die Wichtigkeit dieser Begriffe: Aufgrund der Ambiguität natürlicher Sprachen und deren stark impliziter Bedeutungserzeugung kann ein solcher Schluss immer nur eine gewisse Wahrscheinlichkeit beanspruchen. Tatsächlich beruhen die meisten Verfahren der Textanalyse auf Wahrscheinlichkeitsrechnungen, wodurch sich aber gerade wieder mögliche Verbindungen zu geisteswissenschaftlichen Denkweisen ergeben: Streng genommen treffen digitale Verfahren nämlich gerade keine binären Entscheidungen oder Kategorisierungen, sondern geben statt dessen lediglich die Wahrscheinlichkeiten für Zuordnungen an. Paradoxerweise bieten damit gerade digitale Methoden im Grunde durchaus differenziertere Instrumentarien als bloße Schwarz-Weiß-Entscheidungen (vgl. Craig/Greatly-Hirsch 2017, S. 3). Letztere ergeben sich erst dann, wenn man in die Skala der sich eigentlich erstaunlich analog ausnehmenden Wahrscheinlichkeitswerte nachträglich Schnitte einfügt (etwa einen Schwellwert, ab wann ein Ergebnis nicht als zufällig und daher als hypothesenbelegend anzusehen ist).

Schließlich muss es in Hinblick auf die Anschlussfähigkeit digitaler Methoden auch kein Nachteil sein, dass sich die starke Explizierung und Formalisierung der digitalen Modelle mit den oft komplexen und mehrdeutigen Objekten der Geisteswissenschaften reiben; denn gerade diese Reibung kann sich durchaus erkenntniserweiternd auswirken: Systematisch ließe sich in diesem Sinn etwa zwischen *data-driven* und *data-assisted*

Zugängen unterscheiden (Escobar Varela 2021, S. 7): Während erstere sich auf die strenge Methodik des formalen Modells einlassen und auf replizierbare Ergebnisse abzielen, fragen letztere danach, ob sich bestimmte Phänomene der Quantifizierung widersetzen, und vor allem, warum sie das tun. Der Umgang mit Daten wird hier also zum Ausgangspunkt für ein zwar methodisch geleitetes, aber interpretatives Vorgehen, das digitale Methoden sozusagen ›gegen den Strich‹ liest (vgl. Ramsey 2011).

## 3. Eine digitale Stilgeschichte des Minnesangs?

### 3.1 Minnesangs Meistererzählung

Aufgrund ihrer Detailvergessenheit laufen Untersuchungen aus der Makroperspektive nicht selten Gefahr, in teleologische und nivellierende Meistererzählungen abzugleiten. Dieser Befund gilt jedoch nicht bloß für die digitale Welt, sondern auch für die ›konventionelle‹ geisteswissenschaftliche Forschung. Eine der bekanntesten stilgeschichtlichen Großerzählungen der Altgermanistik ist etwa Hugo Kuhns Erzählung von ›Minnesangs Wende‹ (Kuhn 1952): Nach Kuhn wohnt dem späteren Minnesang eine Tendenz zur Objektivierung inne, die sich daraus ergibt, dass sich die Dichter nicht mehr an der Minne-Konzeption ›subjektiv‹ abarbeiten, sondern diese zur Konvention erstarrt, weshalb sich der Fokus auf die Form verschiebt (Kuhn 1952, S. 143–158; zur Kritik insbesondere Hübner 2013).

Auch in neueren Darstellungen wird dieser Gedanke aufgegriffen: Dass sich die Minnelyrik im 13. Jahrhundert transformiert, wird auch dort nur selten grundsätzlich bestritten (vgl. jedoch Hübner 2013, S. 387–390). Allerdings wird weit stärker die Vielgestaltigkeit und Ausdifferenzierung der Lieder betont, die es – als Einzelfälle – in den Blick zu bekommen gelte. Exemplarisch kann hierfür etwa die Einleitung zum Wolfram-Studien-Band ›Transformationen der Lyrik im 13. Jahrhundert‹ aus dem Jahr 2013

stehen. Dort wird – in deutlicher Anlehnung, aber zugleich Differenzierung der These von Kuhn – als Arbeitsprogramm formuliert:

> Wenn es richtig ist, daß der hochhöfische (›klassische‹) Minnesang innerhalb der Regeln spielt, dagegen der spätmittelalterliche Minnesang mit den Regeln selbst, muß es darauf ankommen, das Verhältnis von Konvention und Abweichung für möglichst viele Parameter zu klären, und zwar zunächst gesondert für jeden überlieferten Einzelfall. (Köbele 2013b, S. 9)

Konzeptionell bedingte Transformation ja, linearer teleologischer Entwicklungszusammenhang nein, so ließe sich also der Wechsel in der Einstellung zur Makroperspektive seit Kuhn resümieren. Für jeden Einzelfall ist gesondert zu eruieren, wie er sich zum Allgemeinen, zur Folie von Konvention und Abweichung, verhält.

### 3.2 Korpora und Wortfrequenzen

Ich möchte im Folgenden zeigen, dass sich dieses Wechselspiel von Verlauf und Einzelfall durchaus auch mit digitalen Mitteln nachzeichnen und zugleich auch als Perspektivenproblem kenntlich machen lässt. Ich nähere mich daher mit unterschiedlichen digitalen Zugängen dieser Frage an. Den Ausgangspunkt soll ein einziges, kurzes Wort und dessen Häufigkeit bilden, nämlich das Wort *ich*, dem eine zentrale Rolle im Minnesang zukommt. Es ist in der Forschung schon öfter bemerkt worden, dass sich die Frage nach der Objektivierung des Sanges über die Betrachtung der Rolle des *Ich* nachzeichnen lässt (vgl. etwa Schnell 2013). Dieser Befund bietet sich nun insbesondere für eine versuchsweise quantitative Auswertung an, da sich die Häufigkeit der Verwendung des Personalpronomens *ich* (und dessen abgeleiteter Formen) leicht auf der Textoberfläche nachzeichnen lässt. Die Frequenz von *ich* bietet also einen guten Indikator für Entwicklungen im Minnesang und ließe sich etwa als zeitliche Verlaufskurve gut darstellen (vgl. hierzu ausführlicher Viehhauser 2017).

Doch offenbart gerade ein solcher, vergleichsweise einfach erscheinender Versuch Probleme, die nun besonders für die spezifischen Situation der mittelalterlichen Literatur virulent werden, denn welches Korpus verwendet man zur Konstruktion einer solchen Verlaufskurve? Ich habe im Folgenden auf Texte zurückgegriffen, die leicht in digitaler Form zugänglich in der Mittelhochdeutschen Begriffsdatenbank (MHDBDB) vorliegen. Dort sind Minnesang-Texte überwiegend aus den klassischen Anthologie-Ausgaben abrufbar. Neben den Liedern aus Minnesangs Frühling (MF) habe ich jene aus Carl von Kraus' Liederdichtern (KLD) und den Schweizer Minnesängern (SM) berücksichtigt. Hinzu kommen die Lieder aus der Walther-Ausgabe (W) von Lachmann/Cormeau (1996) und die Lieder Konrads von Würzburg (KW) aus der Ausgabe Schröder (1924/59). Insgesamt umfasst das Korpus damit 103 Autoren.[3]

Wie leicht ersichtlich ist, werden überlieferungsgeschichtliche Feinheiten damit nicht erfasst. Es gehört zu den großen Leerstellen der digitalen Textanalyse, dass sie für Textvarianten, Mehrfachüberlieferungen oder ähnliches blind bleibt bzw. vielleicht sogar bis zu einem gewissen Grad blind bleiben muss, da sich solche Varianz schwer in die quantitative Analyse einrechnen lässt. Selbst wenn, wie angesichts der Errungenschaften der digitalen Editorik zu hoffen steht, überlieferungsgeschichtlich ausgerichtete Textausgaben in Zukunft auch stärker in digitaler Form zur Verfügung stehen werden, stellt sich weiterhin etwa bei einer Frequenzauszählung die Frage, auf welcher Textbasis diese zu erfolgen hat – aufgrund der Varianz kann der Frequenzbefund für unterschiedliche Textfassungen unterschiedlich ausfallen. Hier kommt also letztlich die oben angesprochene Tendenz der digitalen Textanalyse zur Vernachlässigung von Details zum Tragen: Auf die überlieferungsgeschichtlichen Einzelheiten kommt es nicht an, es zählt der Blick aufs große Ganze. Die gerade durch digitale Zugänge in der Editorik eröffnete Möglichkeit zur Darstellung von Komplexität wird in der Analyse also sogleich wieder nivelliert. Das Problem stellt sich grundsätzlich natürlich auch für neuere Texte, wird

bei mittelhochdeutscher Literatur mit ihrer unfesten Überlieferung aber besonders relevant.

Hinzu kommen Kategorisierungsprobleme: In den Ausgaben von Walther und Konrad etwa findet sich bekanntlich nicht nur Minnesang, sondern auch andere lyrische Formen wie Sangspruch, ohne dass immer eine klare Abgrenzung vorzunehmen ist, und generell sind die Übergänge zwischen den Gattungen fließend. Ich habe mich dafür entschieden, trotz dieser Unsicherheit für die vorliegenden Auswertungen möglichst nur Minnesang zu berücksichtigen. Für die Einschätzung, was genau zum Minnesang dazugehört und was nicht, habe ich mich an den im ›Verfasserlexikon‹ dokumentierten Forschungsstand gehalten und bin den dort verzeichneten Abgrenzungen gefolgt, wohl wissend, dass auch dies eine weitere Reduktion von Genauigkeit im Detail darstellt.

Schließlich stellt sich noch das Problem, wie eine zeitliche Verlaufskurve angesichts der Tatsache, dass sich die meisten Texte nur schwer und unsicher datieren lassen, überhaupt erstellt werden kann. Aus pragmatischen Gründen habe ich mich hier an Autorkorpora und deren gängige, etwa im ›Verfasserlexikon‹ oder bei Hübner 2008 angesetzten Datierungen gehalten. Erneut bringt also die digitale Methode fast notwendigerweise Komplexitätsreduktion mit sich.

Doch wie sieht nun ein solcher natürlich mit Unsicherheit behafteter Überblick über die Geschichte des Minnesangs aus? Für meine erste Analyse habe ich die einzelnen Autorenkorpora in sechs Zeitspannen eingeteilt, und zwar in Kategorie 1, früher Minnesang, 2, hoher Minnesang, sowie 3–6, später Minnesang, den ich wiederum in vier ungefähre zeitliche Phasen eingeteilt habe, die folgende Zeiträume betreffen: SpM 1: Anfang 13. Jahrhundert, SpM 2: Mitte 13. Jahrhundert, SpM 3: Ende 13. Jahrhundert, SpM 4: Ende 13./Anfang 14. Jahrhundert.

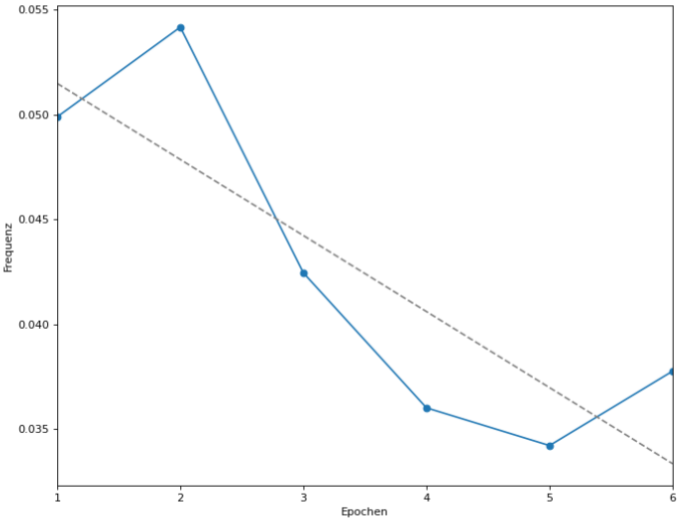


Abb. 1: *ich*-Frequenz im Minnesang nach Epochen

In Abb. 1 sind diese Phasen auf der x-Achse abgebildet, die y-Achse zeichnet die relative Wortfrequenz des Personalpronomens *ich* nach (also die Anzahl der *ich*-Belege geteilt durch die Gesamtanzahl der *tokens* des Teilkorpus). In der Tat zeigen nun die Verlaufskurve und die errechnete Trend-Gerade (gestrichelte Linie) eine Tendenz nach unten, scheinen also die Hypothese des abnehmenden *ich* -Bezugs zu bestätigen.[4]

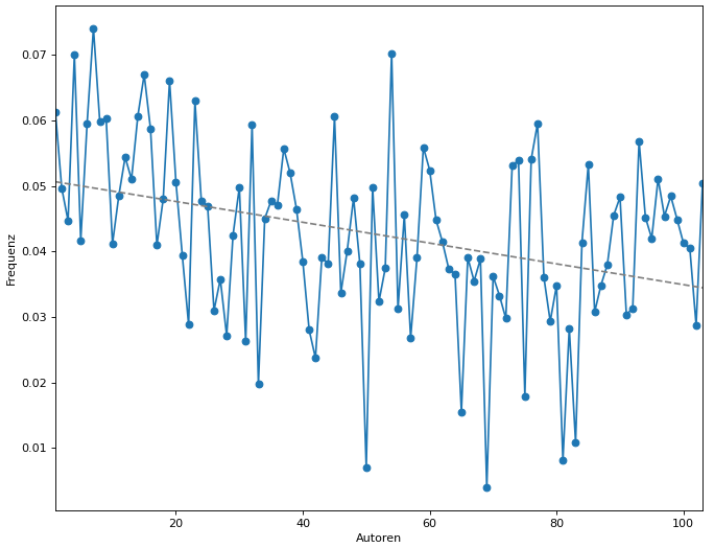


Abb. 2: *ich*-Frequenz im Minnesang nach Autorenkorpora

Ein differenzierteres Bild ergibt sich aber, wenn man, wie in Abb. 2, von der Makroperspektive etwas heranzoomt und nicht die zeitlichen Epochenkorpora, sondern die Autorenkorpora zur Grundlage macht. Hier zeigt sich nun recht augenfällig die oben angesprochene Diversität der Minnesangproduktion: Auch in der Spätphase begegnen durchaus noch Korpora mit durchschnittlichem oder sogar ausgesprochen hohem *ich* -Anteil. Die Frage, ob sich der Minnesang in seiner Ausrichtung grundsätzlich ändert, wird damit nicht zu einer binären Ja-Nein-Frage, sondern letztlich zu einer der Perspektive: Blickt man auf das große Ganze, dann zeigt sich ein Trend, zoomt man in die Details, dann sieht man Diversität, die dem Trend auch zuwiderlaufen kann.

Wie das Beispiel deutlich gemacht hat, lässt sich dieses Wechselspiel nun aber durchaus auch mit quantitativen Methoden fassen, ja sogar, dar-

auf hat So (2017, S. 670) hingewiesen, gerade besonders deutlich und gemäß dem Formalisierungsmodell explizit beschreiben. So könnten etwa die Abweichungen der einzelnen Datenpunkte von der errechneten Trendlinie in den Blick genommen werden: Je weiter sich die Einzelpunkte von der Linie entfernen, desto unsicherer ist es, von einem solchen Trend zu sprechen (So 2017, S. 670). Diese Unsicherheit ist nun aber gerade nicht eine Schwäche des Modells, sondern aus geisteswissenschaftlicher Hinsicht dessen Stärke:

> The advantage of statistical modeling is that it does not present cut- and- dried results that one accepts or rejects. Built into the modeling process is a self-reflexive account of what the model has sought to measure and the limitations of its ability to produce such a measurement. Again, as Box reminds us, »all models are wrong«. What's important is not to insist on how the model is right or nearly right but rather to understand how it is wrong. (So 2017, S. 671)

### 3.3 Hauptkomponentenanalyse (PCA)

Die Geschlossenheit bzw. Diversität der einzelnen Gruppierungen lässt sich auch mit einer etwas komplexeren statistischen Methode als dem bloßen Festhalten von Wortfrequenzen zur Darstellung bringen: Abb. 3 zeigt eine so genannte Hauptkomponentenanalyse bzw. *Principal Component Analysis* (PCA) der häufigsten Wörter (*Most Frequent Words,* abgekürzt MFW) in den zeitlichen Teilkorpora des Minnesangs.[5]

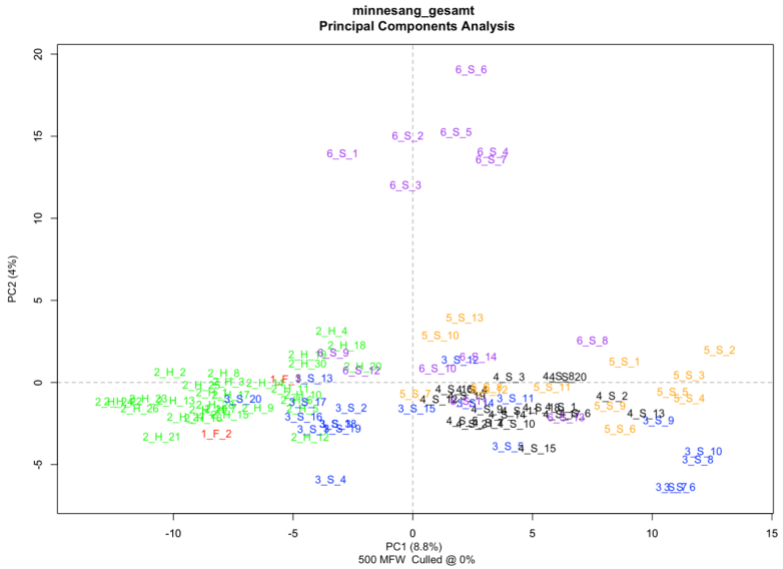


Abb. 3: PCA der Minnesangphasen (unnormalisiert)

In natürlichen Sprachen sind diese MFW üblicherweise Funktionswörter wie Artikel, Konjunktionen oder Präpositionen. In meinem Gesamtkorpus des Minnesangs tritt etwa *ich* mit 8933 Belegen am häufigsten auf; es folgen *daz* (5780), *ir* (4632), *der* (4100), *mir* (3905), *sô* (3485), *und* (2901), *mich* (2641), *mîn* (2624), *si* (2569) und *ist* (2524). Inhaltsbezogenere Wörter begegnen erst weiter hinten in der Liste der MFW; im Minnesang ist das erste dieser Wörter bezeichnenderweise *minne* (an 31. Stelle, mit 1113 Belegen).

Betrachtet man nun die 500 häufigsten Wörter für ein Textkorpus, dann ließe sich deren Frequenzverteilung als Vektor in einem 500-dimensionalen Vektorraum ansetzen, der auf einen Punkt in diesem Vektorraum zeigt. Platziert man in diesem Vektorraum die Wortfrequenz-Vektoren weiterer Korpora, dann lässt sich aus der Entfernung der Vektoren schließen, wie ähnlich die entsprechenden Korpora sind. Im 500-dimensionalen Raum bleiben solche Näheverhältnisse freilich höchst abstrakt. Abhilfe bietet hier

nun das Verfahren der PCA: Bei der PCA werden Variablen, die miteinander korrelieren, möglichst zusammengefasst und damit die Varianz in hochdimensionalen Vektorräumen auf einige wenige Dimensionen, eben die Hauptkomponenten, heruntergerechnet. So werden etwa in einem Datensatz, der verschiedene Schiffe nach deren Wasserverdrängung klassifiziert, die beiden Variablen Länge und Breite vermutlich so stark korrelieren, dass man sie zu einer Komponente (›Größe‹) zusammenfassen könnte.

Üblicherweise werden bei der PCA im Kontext von Textanalysen die ersten beiden Hauptkomponenten der Wortfrequenzverteilung identifiziert, da man diese übersichtlich in einem zweidimensionalen Koordinatensystem eintragen kann. In Abb. 3 habe ich die sechs zeitlichen Teilkorpora in Abschnitte zu jeweils 2000 Wörtern Länge aufgeteilt, die aufgrund ihres Frequenzprofils im zweidimensionalen Raum dargestellt werden können.[6] Abschnitte, die zum selben zeitlichen Teilkorpus gehören, sind durch gleiche Farbe gekennzeichnet (rot: früher Minnesang, grün: hoher Minnesang, blau: später Minnesang 1, schwarz: später Minnesang 2, orange: später Minnesang 3, violett: später Minnesang 4). Die erste Hauptkomponente ist in der Darstellung auf der x-Achse abgebildet, die zweite Hauptkomponente auf der y-Achse. Die Abbildung ist also so zu lesen, dass sich die links platzierten Textabschnitte hinsichtlich der ersten Hauptkomponente stark von jenen unterscheiden, die rechts positioniert sind. Die oben abgebildeten Textabschnitte unterscheiden sich in Hinblick auf die zweite Hauptkomponente von jenen, die unten platziert sind.[7]

Die Darstellung legt zunächst nahe, dass eigentlich nur der hohe Sang (gemeinsam mit den beiden Textabschnitten des frühen Sangs) und der zweite Abschnitt des späten Sangs eine einigermaßen homogene Gruppe ausbilden (in die sich aber auch manche Abschnitte anderer Gruppe einmischen), denn nur bei ihnen liegen die einzelnen Teilabschnitte beieinander, verhalten sich also stilistisch ähnlich. Für diesen Befund sind nun mehrere Deutungen möglich: Er könnte den Umstand reflektieren, dass die zeitliche Zuordnung der Autorenkorpora zu den vier Gruppen des späten

Sangs noch viel unsicherer ist als die der anderen Phasen, oder natürlich schlicht, dass sich eine solche zeitliche Unterscheidung eben nicht in einem klaren stilgeschichtlichen Verlauf abbildet. Vor diesem Hintergrund ist es vielleicht sogar überraschender, dass sich der hohe Sang dann doch vergleichsweise deutlich gruppiert und homogen bleibt, was den Verdacht nährt, dass sich hier ein weiterer Problemfaktor auswirken könnte, der gerade für mittelalterliche deutsche Texte von Belang ist: Nicht zu vergessen ist nämlich, dass die Texte des hohen und frühen Sang aus anderen Textausgaben bezogen sind als der späte Sang (Minnesangs Frühling bzw. die Walther-Ausgabe versus Carl von Kraus' Liederdichter oder die Schweizer Minnesänger). Die (relativ) starke Absonderung könnte also schlicht unterschiedliche Gepflogenheiten der Texteinrichtung in den Editionen reflektieren. Diese Editionsgepflogenheiten setzen wiederum, in mehr oder weniger starker Form, auf den handschriftlichen Quellen mit ihrer nicht regulierten Schreibung auf.

Um dem Zusammenhang eines möglichen Einflusses des Ausgabensignals nachzugehen, habe ich die einzelnen Teilkorpora automatisiert in normalisierte Texte umgewandelt, unterschiedliche Schreibungen also auf eine einheitliche Wortform reguliert. Bis vor kurzem war ein solcher *Preprocessing*-Schritt nur für moderne Texte denkbar, da entsprechende Programme für das Mittelhochdeutsche noch nicht greifbar waren. Mittlerweile liegt aber mit dem von Helmut Schmid entwickelten, auf *Deep-Learning*-Verfahren beruhenden RNNTagger (Recurrent Neural Network Tagger) (Schmid 2019) ein Tool vor, das auch für diese Sprachstufe durchaus brauchbare Ergebnisse liefert. Das Mittelhochdeutsch-Modell des RNNTaggers wurde auf dem Referenzkorpus Mittelhochdeutsch (Klein [u. a.] 2016) trainiert und bietet dementsprechend die Möglichkeit, Texte zu normalisieren, zu lemmatisieren und mit Wortarten-Labels zu versehen (*POS-Tagging*). Zwar können solche Tools niemals absolute Genauigkeit bieten, doch sind die ersten Ergebnisse des Taggers äußerst vielver-

sprechend - und einmal mehr gilt, dass von Unschärfen im Detail für die Makroperspektive zunächst einmal abgesehen werden soll.

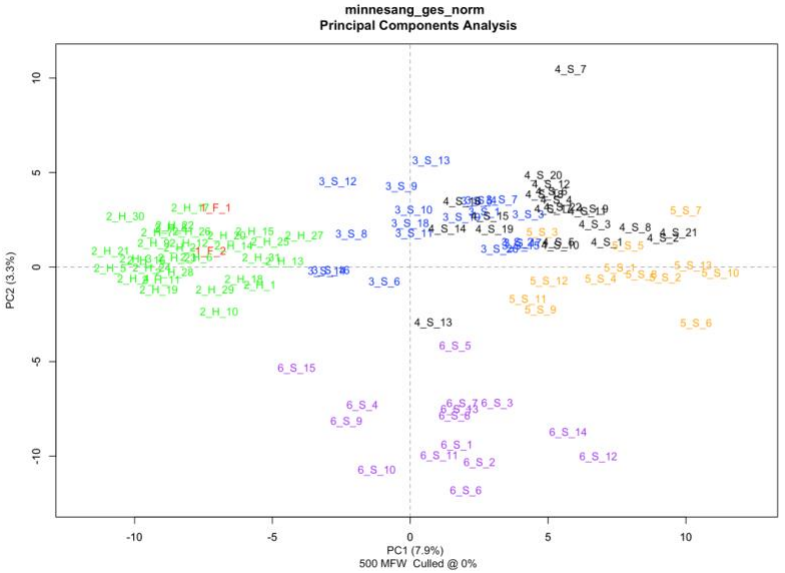


Abb. 4: PCA der Minnesangphasen (normalisiert)

Die PCA der normalisierten Texte (in Abb. 4) zeigt nun, dass sich nach dem *Preprocessing* die Gruppen auch im späten Sang deutlicher abzuzeichnen beginnen.

Nun könnte die Tatsache, dass die Texte als Grundlage für diese Abbildung mit demselben Modell normalisiert worden sind, zur Annahme verleiten, dass sich hier ein genaueres Bild der stilistischen Verhältnisse abzeichnet, da allfällige unterschiedliche Wortformen auf ein und dieselbe Form gebracht werden.

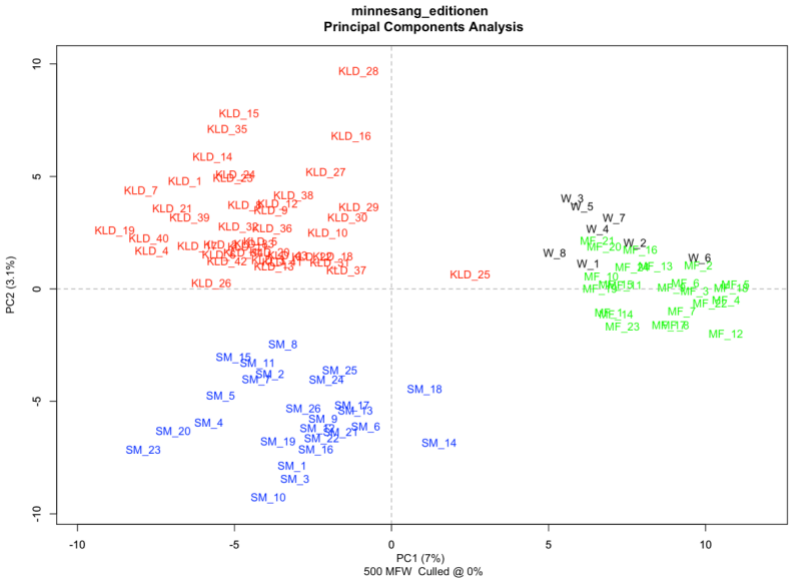


Abb. 5: PCA der Minnesang-Ausgaben (normalisiert)

Doch ist hier Vorsicht geboten, wie Abb. 5 zeigt: Hier habe ich das Gesamtkorpus nicht in zeitliche Unterkorpora aufgeteilt, sondern in solche nach Textausgaben.[8] Und hier wird nun ersichtlich, dass sich die Gruppen, die durch die Ausgaben vorgegeben sind, noch viel klarer abzeichnen (einzig Minnesangs Frühling und die Walther-Ausgabe haben Überschneidungen) und somit, dass das Ausgabensignal also doch eine erhebliche Rolle spielen dürfte.

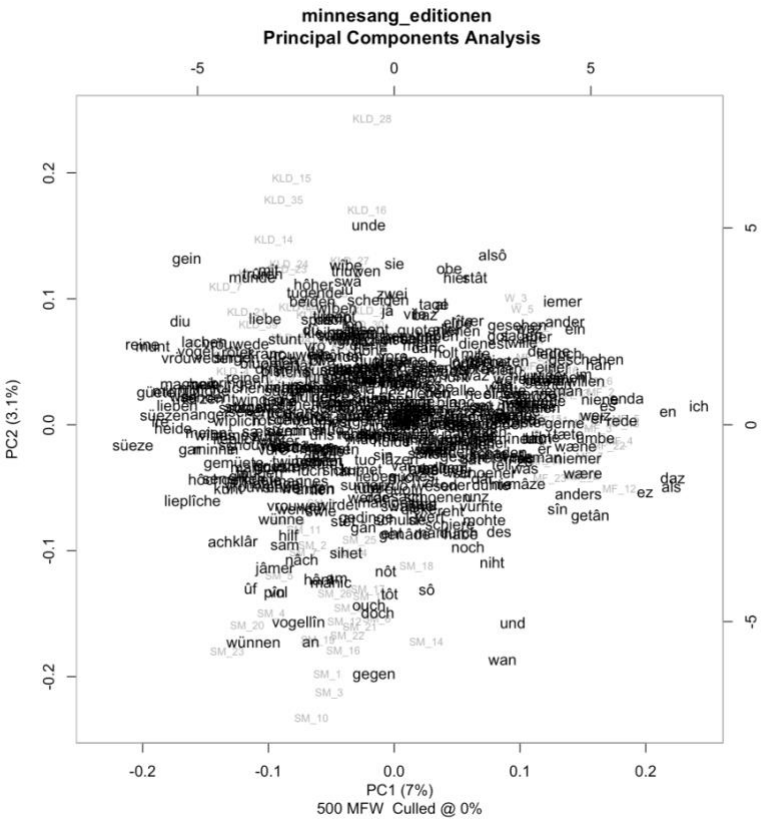


Abb. 6: PCA der Ausgabentexte mit *Loadings* der Wörter

Spätestens hier wäre es nun aufschlussreich zu wissen, welche Wörter für diese Sortierung ausschlaggebend sind. Dies lässt sich mit Hilfe eines sogenannten *Bi-Plots* eruieren, wie er in Abb. 6 ersichtlich ist. In dieser Darstellung wird nicht nur die Verteilung der Texte angezeigt (im Hintergrund in hellgrau), sondern auch die Verteilung der Wörter, die für die Ausprägung der Hauptkomponenten verantwortlich sind. So wird etwa rasch ersichtlich, dass sich die Ausgaben in Bezug auf die y-Achse (also die zweite

Hauptkomponente) z. B. durch die Oppositionen *unde* versus *und* oder *gegen* versus *gein* unterscheiden, die trotz der Normalisierung aufgrund ihrer metrischen Struktur als unterschiedliche Wortformen weitergeführt werden.

Allerdings zeigt die Abbildung auch, dass manche der Wortoppositionen doch auch über bloße Wortvarianten hinausgehen dürften. In Bezug auf die erste Hauptkomponente, die hohen vom späten Sang unterscheidet, nimmt das Wort *Ich* wieder eine prominente Stellung ein, als Gegenpol begegnen auf der rechten Seite Ausdrücke wie *lieplîche*, *süeze* und *süezen*, die sich wohl dem Frauenpreis zuordnen lassen. Dieser gerät also in Opposition zur Ich-Aussage. Ebenfalls auffällig ist die Extremposition von Ausdrücken, die sich dem Natureingangstopos zuordnen lassen (etwa *vogellîn* und *heide*). Wie später noch deutlich wird, stellen diese beiden Gruppen überhaupt ein starkes stilistisches Signal des späten Sangs dar.

### 3.4 Lexikalische Dichte: *Type Token Ratio* (TTR) und *Measure of Textual Lexical Diversity* (MTLD)

Ein weiteres, relativ einfaches Maß zur Beschreibung eines Korpus bietet die *Type-Token-Ratio* (TTR), die das Verhältnis der Gesamtworte (*tokens*) zu den einzelnen Wortformen (*types*) angibt.[9] Dabei kommt insbesondere die Variabilität im Wortschatz in den Blick: Werden in einem Text immer dieselben Wörter verwendet oder viele unterschiedliche Wörter? Der Satz ›ich bin ich‹ würde beispielsweise aus drei *tokens* ›ich‹, ›bin‹, ›ich‹), aber nur zwei *types* bestehen (›ich‹, ›bin‹). Die TTR beträgt demnach zwei Drittel und ist damit geringer als bei dem Satz ›Ich bin Stiller‹, der sich somit als lexikalisch reichhaltiger beschreiben lässt.

Eine unmittelbare Vergleichbarkeit der TTR wird jedoch dadurch erschwert, dass diese nicht unabhängig von der Länge eines Textes ist: Je länger ein Text ist, desto eher wird es vorkommen, dass sich Wörter wiederholen, wodurch die TTR bei längeren Texten automatisch sinkt. Zur Nor-

mierung dieses Textlängenproblems wurden daher auf Basis der TTR weitere, elaboriertere Maßzahlen entwickelt, wie etwa das *Measure of textual lexical diversity* (MTLD, McCarthy 2005). Mit dieser Methode wird die durchschnittliche Länge von Wortsequenzen berechnet, die über einem bestimmten Schwellenwert der TTR liegen. Die Funktionsweise der MTLD besteht also konkret darin, dass der Text Wort für Wort durchgegangen wird und bei jedem Wort die aktuelle TTR berechnet wird. Beim ersten Wort liegt die TTR notwendigerweise noch bei 100 Prozent (das erste *token* muss automatisch auch *type* sein); je weiter man im Text voranschreitet, desto öfter wird es aber vorkommen, dass sich Wörter wiederholen. Passiert diese Wiederholung so oft, dass die TTR unter einen Schwellwert fällt (der normalerweise bei 0,72 liegt), wird der Durchgang durch den Text gestoppt, notiert, wie viele Wörter man vorangekommen ist, und die Zählung von neuem gestartet. Ist der Text durchgearbeitet, kann man dann den Mittelwert der notierten Textsequenzen-Längen berechnen, der die MTLD angibt.

| | Zeit | Autor | Zeichen | Tokens | Types | Mtld | Wortlänge |
|---|---|---|---|---|---|---|---|
| **49** | 4 | Konrad von Wuerzburg | 21289 | 3742 | 1057 | 203 | 5.69 |
| **64** | 5 | Schulmeister von Esslingen | 2134 | 389 | 245 | 197 | 5.49 |
| **68** | 5 | Der Kanzler | 11574 | 2018 | 762 | 185 | 5.74 |
| **82** | 5 | Goeli | 6776 | 1213 | 595 | 177 | 5.59 |
| **94** | 6 | Albrecht Marschall von Raprechtswil | 2660 | 508 | 277 | 175 | 5.24 |
| **6** | 2 | Engelhart von Adelnburg | 869 | 172 | 123 | 173 | 5.05 |
| **101** | 6 | Ulrich von Baumburg | 8370 | 1575 | 654 | 173 | 5.31 |
| **78** | 5 | Konrad von Kirchberg | 6960 | 1279 | 539 | 169 | 5.44 |
| **80** | 5 | Walther von Breisach | 1340 | 243 | 166 | 168 | 5.51 |
| **66** | 5 | Der Duering | 6363 | 1167 | 521 | 164 | 5.45 |

Abb. 7: Statistiken für die zehn Autorenkorpora mit dem höchsten MTLD-Wert

Abb. 7 zeigt nun jene Autorenkorpora im Gesamtkorpus an, die den höchsten MTLD-Wert aufweisen.[10] Besonders hervorstechend ist Konrad von Würzburg mit einer MTLD-Score von 203, es folgen der Schulmeister von Esslingen (197) sowie der Kanzler (185). Dieser Befund ist nun kaum zufällig, sondern hängt vermutlich mit der Tatsache zusammen, dass die genannten Autoren allesamt auch als Verfasser von Sangspruchdichtung in Erscheinung getreten sind.[11] Es scheint sehr wahrscheinlich, dass sich die weniger monothematische Sangspruchproduktion dieser Dichter auch auf ihren Minnesang ausgewirkt hat.[12] Dass für diese eher als ›professionell‹ anzusehenden Dichter andere Maßstäbe gelten, hat auch Hübner (2013, S. 397) konzediert, der letztlich nur für diese Fälle die Kuhn'sche Objektivierungsthese in Ansätzen gelten lassen möchte.

Von den MTLD-Werten schließt sich zudem auffälligerweise wieder der Bogen zu den in Kapitel 3.2 dargestellten *ich*-Frequenzen, denn auch dort wird die Reihenfolge der Autoroeuvres mit der geringsten *ich*-Frequenz vom Kanzler (0.004) und von Konrad von Würzburg (0,007) angeführt, der Schulmeister von Esslingen nimmt Platz fünf dieses Rankings ein (0,015). Zudem finden sich auf den vorderen Plätzen weitere Sangspruch-erprobte Namen wie Walther von Breisach (0,008), aber auch der ausschließlich als Minnesänger bekannte Goeli (0,01), der ebenfalls einen der höchsten MTLD-Werte aufweist (177). Lexikalische Vielfalt und Abkehr vom Ich gehen also hier offensichtlich Hand in Hand und scheinen in vielen Fällen durch den Einfluss der Sangspruch-Form bedingt.

### 3.5 Wordclouds und Distinktivitätsmaße: TF/IDF, *Log Likelihood* und Burrows Delta

Wordclouds bieten eine mittlerweile weit verbreitete Möglichkeit, die MFW eines Korpus zu visualisieren.


Abb. 8: Wordcloud der MFW des Gesamtkorpus

Abb. 8 zeigt etwa eine solche Wordcloud der häufigsten Wörter im Gesamtkorpus des Minnesangs.[13] Die Schriftgröße der Wörter ist dabei in Relation zur Frequenz ausgeführt, je häufiger ein Wort im Korpus auftaucht, desto größer wird es dargestellt.[14] Es zeigt sich hier der oben angesprochene Befund, dass Funktionswörter die häufigsten in einem natürlichsprachlichen Text auftretenden Wörter sind. Im Minnesang dominieren insbesondere das Personalpronomen *ich* und dessen abgeleitete Formen.

Um einen besseren Einblick in den Inhalt der Texte zu haben, lassen sich nun, wie das bei vielen Verfahren der Textanalyse üblich ist, Funktionswörter mit Hilfe von sogenannten Stoppwortlisten aus der Untersuchung ausschließen.


Abb. 9: Wordcloud der MFW des Gesamtkorpus ohne Stoppwörter

Abb. 9 zeigt die Wordcloud der MFW des Gesamtkorpus nach Anwendung einer solchen Stoppwortliste. Übrig bleiben (durchaus erwartungsgemäß) die Leitbegriffe des Minnesangs, *minne*, *herze*, *vröude* und *vrouwe*. Auch ein wenig *leit* und *swaere* ist dabei (aber nicht so ausgeprägt wie die Freude-Komponente). Hinzu kommen schließlich Naturbegriffe (*bluomen, heide*) und Ausdrücke der Sinneswahrnehmung (*ougen*, *sinne*).


Abb. 10: Wordclouds der MFW der Einzelkorpora ohne Stoppwörter

Zoomt man von dieser Makroperspektive einen Schritt hinein und erstellt wie in Abb. 10 Wordclouds für die sechs zeitlichen Teilkorpora, dann ergibt sich ein differenziertes Bild, aus dem sich aber zunächst nur wenig Schlüsse ziehen lassen: Die Leitworte des Minnesangs (*minne, vrouwe, vröude)* dominieren auch hier und bleiben über die ganze Phase des Sangs relevant. Im frühen Sang ist *herze* das häufigste Wort, im hohen Sang *wîp* und in allen Phasen des späten Sangs *minne*. Mehr noch als Veränderung zeigen die *Wordclouds* damit also eine überraschende Konstanz an, Konvention statt Transformation und Wandel.

Eine Möglichkeit, die Unterscheidung der Teilkorpora genauer zu explorieren, bietet das aus dem Kontext des Information Retrieval stammende TF/IDF-Maß, mit dessen Hilfe distinktive Wörter der Subkorpora eruiert werden können (Spärck Jones 1972; zur Anwendung im altgermanistischen Kontext Braun/Reiter 2017). TF steht für *Term Frequency*, IDF für die

*Inverse Document Frequency*. TF/IDF blickt also ebenso wie eine normale Frequenzauszählung zunächst darauf, wie häufig ein Wort vorkommt, setzt dieses Vorkommen aber dazu in Relation, wie ungewöhnlich es ist, dass das Wort öfter auftritt. Dazu wird die TF durch die IDF geteilt, das heißt also, durch die Anzahl der anderen Dokumente aus dem Gesamtkorpus, in der das Wort auch vorkommt. In der TF/IDF-Darstellung (Abb. 11) zeichnen sich nun doch einige zeittypische Wörter ab, z. B. die *merkaere* für den frühen Sang oder der *reien* für die vorletzte Phase des späten Sangs.


Abb. 11: Wordclouds der Teilkorpora gewichtet nach TF/IDF

Schließlich können mit einem weiterem Schritt in das Material hinein die distinktiven Wörter für einzelne Autorenkorpora eruiert werden. Abb. 12–14 zeigen einige Beispiele, bei denen die Distinktivität nicht mit TF/IDF, sondern mit dem *log-likelihood*-Score berechnet wurde.[15] Dieser Score gibt an, ob ein Wort in einem Textkorpus signifikant häufiger erscheint als in den Vergleichskorpora.


Abb. 12: Wordcloud *log likelihood* Wolfram von Eschenbach


Abb. 13: Wordcloud *log likelihood* Reinmar


Abb. 14: Wordcloud *log likelihood* Bernger von Horheim

Bei Wolfram ragt der *urloub* hervor, bei Reinmar die *rede*, bei Bernger von Horheim die *liuge*.

Als nächste Annäherungsstufe könnte man nun in die Einzeltexte selbst hineingehen, und diese qualitativ lesen und interpretieren. Auch wenn der Computer die Makroperspektive betont, braucht man nämlich bei dieser nicht stehenzubleiben. Distant Reading erscheint so nicht als ausschließende Alternative zum Close Reading, sondern die beiden Formen bilden die äußersten Enden eines Kontinuums, zwischen denen man sich in fortwährender Perspektivenverschiebung hin und her bewegen kann. Der Anglist Martin Mueller (2014) hat einen solchen Zugang als Scalable Reading bezeichnet, der also die beiden Extrempositionen der überblicksmäßigen, quantitativen und der detailversessenen, qualitativen Lektüre vereint, und zwar vereint in der Dynamik der Bewegung zwischen den Perspektiven, mithin also in für die Geisteswissenschaften durchaus willkommener Multiperspektivität.

Systematisch gesehen ließen sich TF/IDF und *log likelihood* als Distinktivitätsmaße verstehen, die nicht wie etwa die bloße Auszählung von MFW allgemein auf Wortfrequenzen fokussieren, sondern insbesondere die Unterschiede zwischen zwei Texten bzw. Textgruppen in ihrem Wortgebrauch in den Blick nehmen. Ein weiteres, sehr intuitives und mathematisch relativ einfaches Verfahren stellt Burrows' Zeta dar (Burrows 2007), das anders als die anderen Methoden nicht im Kontext der Computerlinguistik oder des Information Retrieval entwickelt wurde, sondern aus den Digital Humanities selbst stammt (Schöch 2018, S. 81). Zeta zielt nicht auf die absolute Häufigkeit des Wortgebrauchs ab, sondern auf dessen Konsistenz, und zwar auf die Konsistenz des Wortgebrauchs eines Vergleichstextes im Verhältnis zu einem Zieltext (bzw. zwischen entsprechenden Korpora). Das Verfahren ist relativ einfach: Zuerst wird der Vergleichstext in einzelne Abschnitte (etwa zu 2000 *tokens*) zerlegt und dann für jedes Wort ausgezählt, ob und in wie vielen Abschnitten es vorkommt. Dann wird dieselbe Prozedur auf den Zieltext angewendet, und schließlich werden die Werte voneinander abgezogen. Daraus ergibt sich ein Zeta-Score, der zwischen 1 und -1 liegen kann: 1 würde (den in der Praxis natürlich kaum auftretenden Fall) bedeuten, dass das Wort in jedem einzelnen Abschnitt des Vergleichstextes vorkommt und in keinem Abschnitt des Zieltextes. -1 gibt den umgekehrten Fall an. Reiht man die Wörter anschließend nach ihrem Zeta-Score, ergibt sich eine Liste der vom Vergleichstext gegenüber dem Zieltext bevorzugten Wörter, die zugleich die Liste der vom Zieltext vermiedenen Wörter darstellt. Diese sind nur mehr selten Funktionswörter, da deren besonders häufiges Auftreten durch die Abschnittszählung (alle Vorkommnisse zählen pro Abschnitte genau nur wie ein Beleg) abgemildert wird. Damit geraten semantisch leichter zu deutende Wörter aus dem mittleren Frequenzspektrum in den Blick.

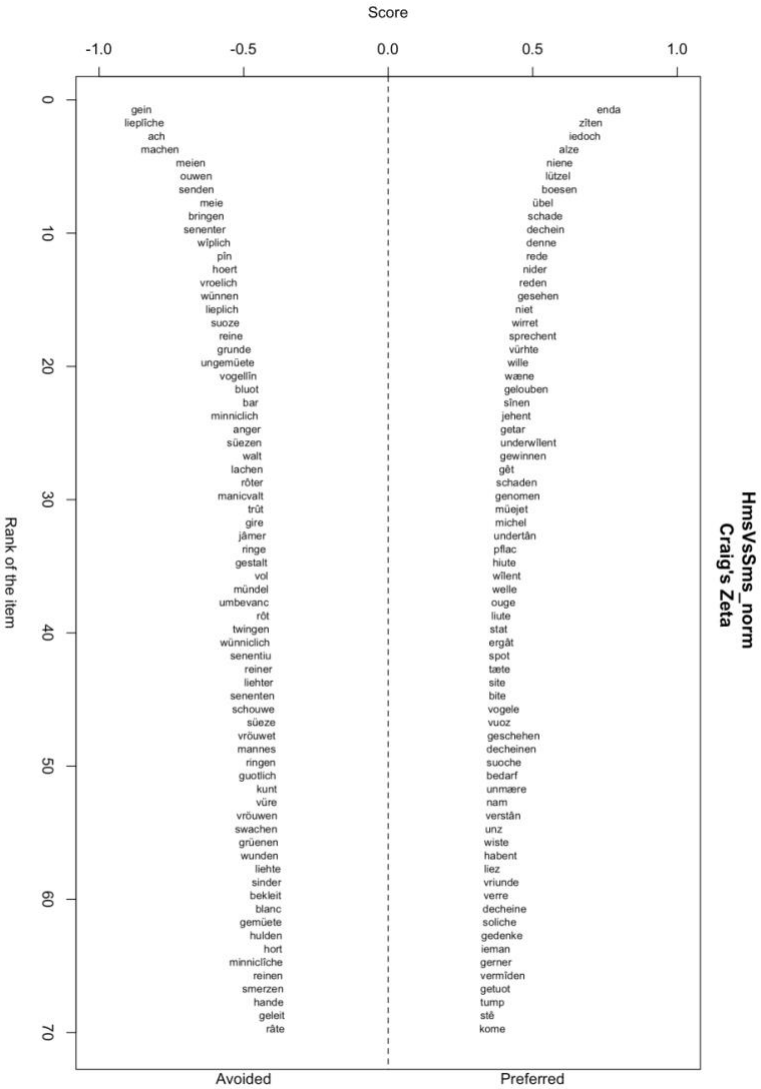


Abb. 15: Bevorzugte bzw. vermiedene Wörter zwischen hohem bzw. frühem Sang und spätem Sang

Zur Erstellung von Abb. 15 habe ich den frühen und hohen Minnesang als Vergleichstextkorpus den vier Teilkorpora des späten Sangs gegenübergestellt.[16] Rechts befinden sich die von der ersten Phase des Minnesangs bevorzugten Wörter, links die vermiedenen. Der Zeta-Score ist dabei auf der x-Achse abgetragen. Je weiter ein Wort also von der Mittellinie entfernt ist, desto deutlicher gehört es zu einer der beiden Gruppen. Die y-Achse gibt das Ranking der Wörter wieder, die einen Zeta-Score von 0,3 über- bzw -0,3 unterschreiten.

Die genauere Durchsicht bringt auf der Seite der in der Früh- und Hochphase vermiedenen (und damit in der Spätphase bevorzugten) Wörter den einheitlicheren Befund: Immer wieder begegnen hier Wörter, die mit dem Natureingangstopos in Verbindung zu bringen sind (*meie, meien, ouwen, vogellîn*), der sich ja schon in anderen Analysen als typisches Formmerkmal der Spätphase erwiesen hat.[17] Wörter wie *wîplich, lieplich, suoze, reine* deuten auf den Frauenpreis hin. Einen interessanten Einzelbefund stellt die Spitzenstellung der Interjektion *ach* dar, die offenkundig im späteren Sang häufiger auftritt. Als Wort mit dem ausgeprägtesten negativen Zeta-Score begegnet aber das schon bei der PCA der nach Editionen geordneten Texte zu Tage getretene einsilbige *gein*. Die vom Vergleichskorpus bevorzugten Wörter präsentieren sich deutlich uneinheitlicher. Auch hier steht an der Spitze mit *enda* vermutlich ein Ausgaben-Artefakt, danach begegnen vermehrt negative Wörter (*boesen, übel, schade*), auch das Wortfeld *rede*, *reden* und *sprechent* wird in den früheren Texten konsistent häufiger verwendet.

### 3.7 Distributionelle Semantik und Topic Models

Das wohl am ehesten geeignete Framework, um mit dem Computer über die reine Textoberfläche hinaus in tiefere Bedeutungsschichten vorzudringen, bietet das Theoriegebäude der distributionellen Semantik. Die distributionelle Semantik geht davon aus, dass sich die Bedeutung eines Wortes

nicht (oder nicht nur) aus sich selbst ergibt, sondern aus dem Kontext, in dem es erscheint. Der Linguist John Rupert Firth hat diese Grundannahme in dem berühmten Zitat »you shall know a word by the company it keeps« zusammengefasst (Firth 1957, S. 11). Dieser Umstand bietet nun auch dem Computer die Chance, komplexere und auch implizite Bedeutungen zu erkennen. So wird es etwa gemäß der distributionellen Hypothese für den Computer möglich, zwischen unterschiedlichen Bedeutungen des oberflächlich gleichen Wortes ›Bank‹ zu differenzieren: Begegnet dieses überwiegend im Kontext von anderen finanzbezogenen Ausdrücken, so ist der Bedeutungsaspekt von Bank als Geldinstitut wahrscheinlich, begegnen hingegen im Umfeld Naturausdrücke, so ist die Bedeutung als Parkbank eher anzunehmen. Dieses Beispiel zeigt bereits, dass auch die distributionelle Semantik nicht hundertprozentig sichere Ergebnisse liefern kann, sondern mit mehr oder minder großen Wahrscheinlichkeiten zu rechnen hat.

Eines der bekanntesten Verfahren, das auf den Grundsätzen der distributionellen Semantik beruht, ist das Topic Modeling, mit dessen Hilfe die Themenstruktur von größeren Korpora nachgezeichnet werden kann (Blei [u. a.] 2003; ausführlich zum Verfahren Horstmann 2018). Beim Topic Modeling wird von dem gemeinsamen Auftreten bestimmter Wörter auf zugrundeliegende Themen- oder besser gesagt *topic*-Cluster geschlossen.[18] So deutet etwa der Umstand, dass in einem Text sehr oft die Wörter ›Fisch‹, ›Boot‹, ›Netz‹ gemeinsam auftreten, darauf hin, dass es in diesem Text um ein gemeinsames Thema geht, das sich als ›Fischerei‹ benennen lässt. Ein Text kann und wird dabei durchaus mehrere solcher *topics* (ausgeprägt) aufweisen; ebenso kann ein Wort in unterschiedlichen *topics* (prominent) auftreten.[19]

Beim Topic Modeling handelt es sich um ein unüberwachtes Machine-Learning-Verfahren, das heißt, der Computer fügt in mehreren Trainingsdurchgängen, bei der die Kontextwahrscheinlichkeiten immer besser eingeschätzt werden, selbständig bestimmte Wörter zu Themenclustern zusammen. Wie viele solcher Themencluster angesetzt werden sollen, muss

jedoch von der menschlichen Benutzer*in vorgegeben werden; auch die Interpretation und letztendliche Benennung der *topics* bleibt dem Menschen überlassen. Der Computer stellt also nur fest, welche Wörter überzufällig oft gemeinsam auftreten, was auf eine latente Variable, nämlich einen semantischen Zusammenhang in einem *topic*, hinweisen könnte.


Abb. 16: Wordclouds der *topics* des Minnesangkorpus

Abb. 16 zeigt ein solches Topic Model des Minnesangs.[20] Ich habe vom Computer 15 *topics* berechnen lassen. Die einzelnen Topics sind in Wordcloud-Darstellungen visualisiert, wobei hier nun in proportionaler Größe jene Wörter angezeigt werden, die am meisten zur Formierung eines *topics* beitragen (und damit für das Thema am bedeutendsten sind). Neben erwartbaren Minne-spezifischen Themenclustern zeigen sich im Topic Model einige interessante Einzelcluster, etwa Topic Nr. 14, das mit Wörtern wie *sprach, liebe, tac, scheiden, naht, klage, wahtaer* und *ritaer* die Konstellation des Tageliedes reflektiert (vierte Wortcloud in der letzten Reihe) oder *topic* Nr. 5, das mit Wörtern wie *bluomen, winter, meien, vogel, heide, sumer* dem Natureingangstopos entspricht.

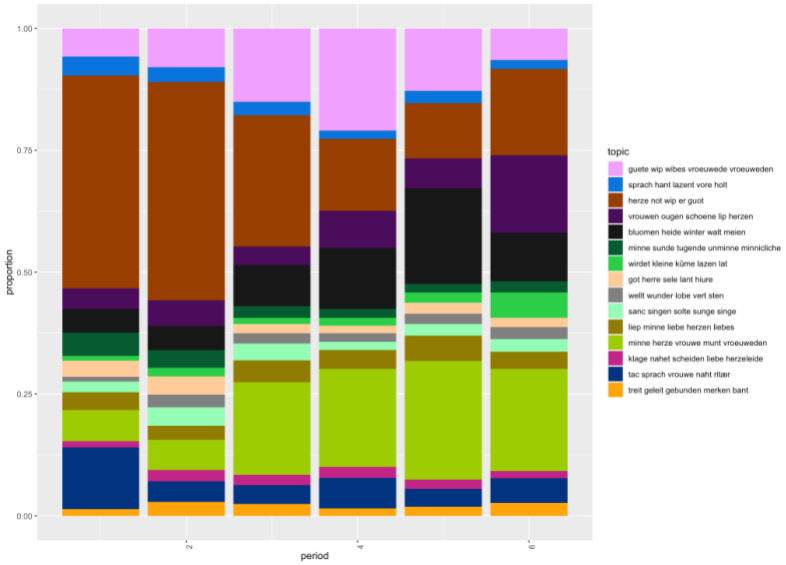


Abb. 17: Zeitliche Verteilung der *topics* im Minnesangkorpus

Auch im Kontext der Topic-Model-Analyse lässt sich demonstrieren, dass der Natureingang für den späten Sang kennzeichnend ist. Abb. 17 zeigt den zeitlichen Verlauf der Konjunktur der fünfzehn *topics*, die die Maschine berechnet hat. Die sechs Spalten auf der x-Achse stehen für die sechs zeitlichen Teilkorpora ein, auf der y-Achse repräsentieren die unterschiedlichen Farben den Anteil der 15 Topics an den Texten der Zeitstufe. Je breiter das Band eines *topics* ausfällt, desto mehr Textabschnitte lassen sich ihm mit höherer Wahrscheinlichkeit zuordnen. Und hier zeigt sich nun, dass das in Schwarz dargestellte *topic* 2, das den Natureingang abbildet, im Verlauf des späten Sangs immer mehr zunimmt und im fünften Abschnitt seinen Höhepunkt erreicht.

Andere, sich ebenfalls stark verändernde *topic*-Bänder sind weniger einfach zu deuten. So ist die Frühzeit etwa durch einen hohen Anteil von *topic* 3 geprägt, in dem die Wörter *herze, not, wip, êr* und *guot* dominieren. Daneben zeichnen sich jedoch Begriffsfelder der Klage (*leit*) ebenso wie der

*vröuwede* ab. Das *topic* scheint im weiteren Verlauf der Gattung durch *topic* 13 (*minne, herze, vrouwe, munt, vrouwe*) abgelöst zu werden.

## 4. Fazit

Mit meinem Durchgang durch einige Methoden der digitalen Textanalyse wollte ich demonstrieren, dass diese zwar notwendigerweise zur Darstellung größerer Zusammenhänge tendieren und damit Details nivellieren, dass man gerade mit digitalen Methoden aber auch die Perspektivenabhängigkeit dieser großen Zusammenhänge herausstellen kann. So lässt sich zwar die ›Geschichte‹ des späten Minnesangs als Transformation beschreiben, zugleich aber auf die immer noch gegebene Vielfältigkeit der stilistischen Muster verweisen.

Dabei haben sich deutlich spezifische Probleme gezeigt, die sich gerade bei der Anwendung digitaler Methoden auf mittelalterliche volkssprachige Literatur ergeben. Zwar ist es mittlerweile möglich, durch die Entwicklung des RNNTaggers Texte zu normalisieren und damit den ›Störfaktor‹ der uneinheitlichen Schreibung des Mittelhochdeutschen abzumildern. Doch wirkt die ungeregelte Orthographie, wie es scheint, trotzdem noch nach, etwa wenn unterschiedliche Textausgaben unterschiedliche Wortformen präferieren oder sich diese aus metrischen Gegebenheiten ergeben. Wie die PCA der nach Ausgaben gegliederten Texte gezeigt hat, dürften Störfaktoren im stilistischen Signal, die über eine ›normale‹ Stilmischung moderner Texte hinausreichen, nicht zu vernachlässigen sein.

Zudem stellt sich das Problem der Kategorienbildung: Zuweisungen von Texten zu Autorkorpora und deren Datierung sind auf unsicherem Boden gebaut. Dennoch bieten sie einen Ausgangspunkt, dessen Unzuverlässigkeit gerade durch digitale Methoden (wie etwa mit Hilfe der Herausstellung einer uneinheitlichen Verteilung der Textpunkte in der PCA) ausgestellt werden kann und unbedingt auch werden soll: Gerade die Abweichungen

von der Idealvorstellung des statistischen Modells sind besonders aufschlussreich.

Blickt man auf den inhaltlichen Ertrag, den die Analysen erbracht haben, dann zeigen sich immer wieder Einzelergebnisse, die vielleicht nicht grundsätzlich neu sind (das Korpus des Minnesangs, das hier exemplarisch herangezogen wurde, ist letztlich nicht so groß, dass es nicht auch qualitativ zu überschauen wäre), aber durch den stilistischen Befund bestehende Vermutungen bestärken können, neue Anregungen bieten und möglicherweise Offensichtliches, aber doch bislang Übersehenes zu Tage bringen. So konnte die Frequenzanalyse in Zusammenspiel mit der Untersuchung der lexikalischen Dichte den Verdacht erhärten, dass Transformationen des Minnesangs insbesondere bei Autoren auftreten, die auch als Sangspruchdichter in Erscheinung getreten sind und vermutlich von dort neue Formen in den Minnesang eingebracht haben. Es wurde zudem deutlich, wie stark der Natureingang als stilistisch-thematisches Spezifikum des späten Sanges in Erscheinung tritt. Und schließlich bieten die gezeigten Verfahren Anknüpfungspunkte für ein Scalable Reading der Texte, das von den Befunden auf der quantitativen Makro-Ebene seinen Ausgang nehmen und von dort auch wieder auf die qualitative Mikro-Ebene zurückführen kann.

Digitale Methoden sind so gesehen nicht der Endpunkt der Interpretation, sondern im Gegenteil erst der Ausgangspunkt; sie liefern keine endgültige Ergebnisse, die ohne Weiteres hinzunehmen sind, sondern Anregungen, die sich gerade aus der Friktion der digitalen Modelle mit den analogen Gegenständen ergeben können.

## Anhang A: Verzeichnis der erfassten Autorenkorpora

| ID | Abschnitt | Autor | Lieder | Zeichen | Tokens | Types | Ausgabe |
|---|---|---|---|---|---|---|---|
| **1** | 1 | Der von Kürenberg | 2 | 2831 | 556 | 286 | MF |
| **2** | 1 | Burggraf von Regensburg | 2 | 876 | 171 | 109 | MF |
| **3** | 1 | Dietmar von Eist | 16 | 10785 | 2097 | 640 | MF |
| **4** | 1 | Burggraf von Rietenburg | 2 | 1844 | 361 | 194 | MF |
| **5** | 1 | Meinloh von Sevelingen | 3 | 4198 | 799 | 351 | MF |
| **6** | 1 | Kaiser Heinrich | 3 | 2071 | 409 | 218 | MF |
| **7** | 2 | Engelhart von Adelnburg | 2 | 869 | 172 | 123 | MF |
| **8** | 2 | Ulrich von Gutenburg | 4 | 12382 | 2468 | 695 | MF |
| **9** | 2 | Friedrich von Hausen | 17 | 15594 | 3122 | 747 | MF |
| **10** | 2 | Heinrich von Veldeke | 37 | 15068 | 2922 | 842 | MF |
| **11** | 2 | Walther von der Vogelweide | 76 | 91975 | 17857 | 2485 | WL |
| **12** | 2 | Rudolf von Fenis | 8 | 8193 | 1627 | 493 | MF |
| **13** | 2 | Albrecht von Johansdorf | 13 | 12528 | 2431 | 728 | MF |
| **14** | 2 | Heinrich von Rugge | 12 | 14158 | 2770 | 727 | MF |
| **15** | 2 | Bernger von Horheim | 6 | 5165 | 1023 | 376 | MF |
| **16** | 2 | Hartwig von Rute | 4 | 2020 | 400 | 207 | MF |
| **17** | 2 | Bligger von Steinach | 2 | 1486 | 299 | 178 | MF |
| **18** | 2 | Heinrich von Morungen | 35 | 31605 | 6231 | 1285 | MF |
| **19** | 2 | Reinmar | 70 | 79610 | 15866 | 1909 | MF |
| **20** | 2 | Hartmann von Aue | 18 | 18407 | 3631 | 882 | MF |
| **21** | 2 | Gottfried von Straßburg | 2 | 2893 | 544 | 300 | MF |
| **22** | 2 | Wolfram von Eschenbach | 9 | 10022 | 1898 | 699 | MF |
| **23** | 3 | Hiltbolt von Schwangau | 22 | 13014 | 2609 | 644 | KLD |
| **24** | 3 | Otto von Botenlauben | 12 | 6885 | 1357 | 500 | KLD |
| **25** | 3 | Rubin | 21 | 23699 | 4556 | 980 | KLD |
| **26** | 3 | Der tugendhafte Schreiber | 11 | 10951 | 2059 | 684 | KLD |
| **27** | 3 | Gottfried von Neifen | 51 | 56230 | 10449 | 1461 | KLD |
| **28** | 3 | Burkhard von Hohenvels | 18 | 21670 | 4005 | 1133 | KLD |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **29** | 3 | Der Markgraf von Hohenburg | 6 | 4567 | 924 | 322 | KLD |
| **30** | 3 | Wachsmuot von Künzingen | 7 | 5673 | 1100 | 420 | KLD |
| **31** | 3 | Christan von Hamle | 6 | 5846 | 1075 | 477 | KLD |
| **32** | 3 | Friedrich der Knecht | 5 | 6128 | 1191 | 449 | KLD |
| **33** | 3 | Friedrich von Leiningen | 1 | 1575 | 308 | 189 | KLD |
| **34** | 3 | Heinrich von Anhalt | 2 | 1665 | 323 | 190 | KLD |
| **35** | 3 | Rudolf von Rotenburg | 11 | 11071 | 2127 | 618 | KLD |
| **36** | 3 | Ulrich von Munegiur | 3 | 2467 | 494 | 237 | KLD |
| **37** | 3 | Walther von Mezze | 10 | 11060 | 2152 | 679 | KLD |
| **38** | 3 | Hesso von Rinach | 2 | 1349 | 254 | 166 | SM |
| **39** | 3 | Ulrich von Singenberg | 31 | 33457 | 6445 | 1244 | SM |
| **40** | 4 | Markgraf Heinrich von Meißen | 6 | 4455 | 843 | 373 | KLD |
| **41** | 4 | Hugo von Werbenwag | 5 | 4149 | 763 | 368 | KLD |
| **42** | 4 | Herrand von Wildonie | 3 | 2200 | 431 | 231 | KLD |
| **43** | 4 | Der Kol von Niunzen | 4 | 1475 | 285 | 174 | KLD |
| **44** | 4 | Reinmar von Brennenberg | 5 | 12028 | 2243 | 717 | KLD |
| **45** | 4 | Der Schenk von Limburg | 6 | 6545 | 1235 | 462 | KLD |
| **46** | 4 | Ulrich von Liechtenstein | 59 | 74826 | 14224 | 1869 | KLD |
| **47** | 4 | Ulrich von Winterstetten | 40 | 54613 | 10236 | 1643 | KLD |
| **48** | 4 | Bruno von Hornberg | 4 | 3507 | 680 | 313 | KLD |
| **49** | 4 | Burggraf von Lienz | 2 | 2633 | 507 | 263 | KLD |
| **50** | 4 | Konrad von Würzburg | 23 | 21289 | 3742 | 1057 | KW |
| **51** | 4 | Der von Sachsendorf | 7 | 5965 | 1147 | 433 | KLD |
| **52** | 4 | Wachsmuot von Mühlhausen | 5 | 2830 | 535 | 271 | KLD |
| **53** | 4 | Waltram von Gresten | 3 | 1787 | 359 | 197 | KLD |
| **54** | 4 | Willehelm von Heinzenburg | 5 | 2931 | 569 | 286 | KLD |
| **55** | 4 | Der von Stadegge | 3 | 2057 | 392 | 218 | KLD |
| **56** | 4 | Der von Suonegge | 3 | 2145 | 401 | 199 | KLD |
| **57** | 4 | Der von Wissenlo | 4 | 2704 | 528 | 249 | KLD |
| **58** | 4 | Günther von dem Forste | 6 | 9966 | 1909 | 538 | KLD |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **59** | 4 | Heinrich von der Mure | 3 | 2266 | 463 | 232 | KLD |
| **60** | 4 | König Konrad der Junge | 2 | 1323 | 250 | 147 | KLD |
| **61** | 4 | Rudolf der Schreiber | 3 | 3633 | 701 | 327 | KLD |
| **62** | 4 | Heinrich von Sax | 4 | 5272 | 1004 | 395 | SM |
| **63** | 4 | Walther von Klingen | 8 | 7402 | 1362 | 494 | SM |
| **64** | 5 | Der wilde Alexander | 2 | 2522 | 480 | 245 | KLD |
| **65** | 5 | Schulmeister von Esslingen | 2 | 2134 | 389 | 245 | KLD |
| **66** | 5 | Brunwart von Augheim | 5 | 3453 | 654 | 294 | KLD |
| **67** | 5 | Der Düring | 7 | 6363 | 1167 | 521 | KLD |
| **68** | 5 | Der Dürner | 1 | 1489 | 289 | 178 | KLD |
| **69** | 5 | Der Kanzler | 12 | 11574 | 2018 | 762 | KLD |
| **70** | 5 | Der Püller | 5 | 4196 | 777 | 346 | KLD |
| **71** | 5 | Konrad von Landeck | 22 | 32201 | 5942 | 1195 | SM |
| **72** | 5 | Der von Buchein | 3 | 1643 | 306 | 183 | KLD |
| **73** | 5 | Der von Obernburg | 7 | 5812 | 1104 | 410 | KLD |
| **74** | 5 | Der von Scharpfenberg | 2 | 2298 | 455 | 237 | KLD |
| **75** | 5 | Der von Stammheim | 1 | 3271 | 633 | 322 | KLD |
| **76** | 5 | Hartmann von Starkenberg | 3 | 1557 | 306 | 177 | KLD |
| **77** | 5 | Herzog Heinrich von Breslau | 2 | 2584 | 493 | 252 | KLD |
| **78** | 5 | König Wenzel von Böhmen | 3 | 4379 | 842 | 376 | KLD |
| **79** | 5 | Konrad von Kirchberg | 6 | 6960 | 1279 | 539 | KLD |
| **80** | 5 | Markgraf Otto von Brandenburg | 7 | 4088 | 752 | 371 | KLD |
| **81** | 5 | Walther von Breisach | 1 | 1340 | 243 | 166 | KLD |
| **82** | 5 | Der Taler | 3 | 3543 | 689 | 326 | SM |
| **83** | 5 | Goeli | 4 | 6776 | 1213 | 595 | SM |
| **84** | 5 | Heinrich von Frauenberg | 5 | 3863 | 743 | 322 | SM |
| **85** | 5 | Heinrich von Stretelingen | 3 | 2658 | 496 | 232 | SM |
| **86** | 5 | Heinrich von Tettingen | 2 | 1749 | 331 | 180 | SM |
| **87** | 5 | Konrad von Altstetten | 3 | 2845 | 553 | 280 | SM |
| **88** | 5 | Kraft von Toggenburg | 7 | 7487 | 1413 | 474 | SM |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **89** | 5 | Steinmar | 14 | 14868 | 2818 | 801 | SM |
| **90** | 5 | Winli | 8 | 7500 | 1436 | 513 | SM |
| **91** | 6 | Johannes Hadlaub | 54 | 70350 | 13898 | 2022 | SM |
| **92** | 6 | Christian von Lupin | 7 | 5484 | 1072 | 439 | KLD |
| **93** | 6 | Gösli von Ehenheim | 2 | 2126 | 391 | 222 | KLD |
| **94** | 6 | Heinrich Hetzbold von Weißensee | 8 | 5697 | 1095 | 416 | KLD |
| **95** | 6 | Albrecht Marschall von Raprechtswil | 3 | 2660 | 508 | 277 | SM |
| **96** | 6 | Der von Gliers | 3 | 15153 | 2995 | 825 | SM |
| **97** | 6 | Der von Trostberg | 6 | 5217 | 967 | 432 | SM |
| **98** | 6 | Heinrich Rost zu Sarnen | 9 | 7639 | 1407 | 557 | SM |
| **99** | 6 | Heinrich Teschler | 13 | 15788 | 3033 | 851 | SM |
| **100** | 6 | Jakob von Warte | 6 | 7541 | 1403 | 479 | SM |
| **101** | 6 | Otto zum Turm | 5 | 5933 | 1104 | 455 | SM |
| **102** | 6 | Ulrich von Baumburg | 7 | 8370 | 1575 | 654 | SM |
| **103** | 6 | Wernher von Hohenberg | 8 | 5068 | 1010 | 396 | SM |

Abschnitte:

1 – Früher Minnesang

2 – Hoher Minnesang

3 – Später Minnesang 1 (Anfang 13. Jh.)

4 – Später Minnesang 2 (Mitte 13. Jh.)

5 – Später Minnesang 3 (Ende 13. Jh.)

6 – Später Minnesang 5 (Ende 13. / Anfang 14. Jh.)

## Anmerkungen

1 Vgl. hierzu die klassische Beschreibung der Grenze zwischen nomothetischen und idiographischen Wissenschaften bei Windelband 1910.

2 Schon durch diese Explizierung können sich wissenschaftliche Mehrwerte ergeben, vgl. dazu Gius/Jacke 2015.

3 Es versteht sich von selbst, dass durch eine solche Auswahl keine Vollständigkeit zu erreichen ist, wie sie etwa Moretti vorgeschwebt haben mag (zur Frage, wa-

rum eine solche Vollständigkeit ohnedies Chimäre bleibt, vgl. Rosen 2011). Da die unterschiedliche Gestaltung von Textausgaben nicht unerheblichen Einfluss auf textanalytische Methoden hat (siehe dazu unten), habe ich versucht, die Zahl der Ausgaben einigermaßen klein zu halten, und dies bei größtmöglicher Abdeckung der Minnesangproduktion (eine größere Leerstelle stellt jedenfalls der überlieferungsgeschichtlich schwierige Neidhart dar, sonst ist ein Großteil der Minnesangproduktion vertreten). Eine genaue Aufstellung der berücksichtigten Autoren findet sich im Anhang.

4 Die Trendlinie wird mit der Methode der kleinsten Quadrate mithilfe der Funktion linregress des Python-Scipy-Packages berechnet. Ich habe dafür auf den Code zurückgegriffen, der in Karsdorp [u. a.] 2021, S. 21, beschrieben ist.

5 Die Darstellung wurde mit dem R-Package *stylo* von Eder [u. a] 2013 erstellt (Parameter: 500 MFW, Sample-Größe 2000 Wörter, correlation PCA). Zum Verfahren und dessen Anwendung in der Stilistik vgl. Craig/Greatley-Hirsch 2017.

6 Die Punkte werden dabei durch entsprechende Namenskürzel angezeigt: »3_S_13« gibt also z. B. den Punkt für den 13. Abschnitt von 2000 Wörtern im dritten der sechs Teilkorpora an; es handelt sich also um Lyrik vom Anfang des 13. Jahrhunderts. Das »S« zeigt an, dass sich das Teilkorpus dem späten Sang zuordnen lässt, 3_S wäre also das erste Teilkorpus des späten Sangs.

7 Zu beachten ist bei diesem und den folgenden Beispielen, dass die Hauptkomponenten im Fall des Minnesangkorpus recht schwach ausgeprägt sind. Die erste Hauptkomponente steht in Abb. 3 gerade einmal für 8,8% der Varianz ein, deckt also nur weniger als ein Zehntel der Gesamtvarianz ab. In den folgenden Beispielen ist der Wert noch geringer.

8 Die Konrad von Würzburg-Ausgabe habe ich dabei aufgrund ihrer Kürze (im Vergleich zu den anderen Anthologien) und der Besonderheiten des Konrad-Lied-Korpus beiseite gelassen.

9 Einen Überblick über die angeführten Methoden bieten Perkhun [u. a.] 2011

10 Die Berechnung erfolgte mit dem Python-Package *LexicalRichness* (Shen 2022) auf dem normalisierten Korpus.

11 Ich danke Sonja Glauch für entsprechende Hinweise. Bei Konrad ist allerdings wieder das Ausgaben-Problem zu berücksichtigen, da seine Lieder ja als einzige aus einer eigenen Edition bezogen sind.

12 Dass der TTR-Wert für Sangspruch höher liegt als der für den Minnesang bestätigen auch Braun/Reiter 2017, S. 15.

13 Die Wordclouds wurden mit dem Python-*wordcloud*-Package erstellt, unter Rückgriff auf den Code von https://towardsdatascience.com/how-to-make-word-clouds-in-python-that-dont-suck-86518cdcb61f.

14 Freilich können Wordcloud-Darstellungen auch trügerisch sein, da z. B. längere Wörter per se größer erscheinen als kürzere Wörter und dadurch überbetont werden. Doch gilt hier das oben für das Modell Gesagte: Daten-Visualisierungen sollten keinesfalls mit der Wahrheit ›an sich‹ verwechselt werden, sondern Ausgangspunkte für Interpretationen bieten.

15 Die Erstellung der Wordclouds erfolgte mit dem *wordclouds*-Package in R (Fellows 2018), unter Rückgriff auf Code von Wiedeman/Niekler 2017.

16 Die Grafik wurde mit der oppose()-Funktion es *stylo*-Package für R erstellt (Eder [u. a.] 2013), auf der Basis von Textabschnitten zu 3000 Wörtern.

17 Siehe dazu die Ergebnisse der PCA sowie schon Schnell 2013, S. 326, der so weit geht, dass er den Natureingang als neues Gattungssignal des Minnesangs im Spätmittelalter ansieht, das notwendig werde, da Sangspruch und Minnesang immer ununterscheidbarer werden.

18 Eine genaue Übersetzung der *topics* mit Themen wäre allerdings irreführend: *topics* können sich auch durch andere Bedeutungsformationen als Themen ergeben, etwa durch Einsprengsel von fremdsprachigen Ausdrücken, aber auch durch die Zugehörigkeit von Wörtern zu allgemeineren Gruppierungen wie etwa bei Ausdrücken der Zeit (vgl. Schöch 2017, S. 4)

19 Streng genommen sind sogar alle *topics* in einem Text enthalten und alle Wörter in allen *topics*, manche jedoch mit nur sehr geringer Wahrscheinlichkeit.

20 Erstellt in Anlehnung an Wiedeman/Niekler 2017 mit dem Package *topicmodels* in R (Grün/Hornik 2011). Siehe zum Topic Model des Minnesangs ausführlicher Viehhauser 2017. Im Gegensatz zur Darstellung dort wurde das Modell auf Grundlage der mit dem RNNTagger normalisierten Texte erstellt.

## Literaturverzeichnis

### Primärliteratur

KLD - Deutsche Liederdichter des 13. Jahrhunderts, hrsg. von Carl von Kraus, Tübingen 1952.

KW - Kleinere Dichtungen Konrads von Würzburg, hrsg. von Edward Schröder mit einem Nachwort von Ludwig Wolff, 3. Aufl, Berlin 1924/59.

MF - Des Minnesangs Frühling, unter Benutzung der Ausgaben von Karl Lachmann/Moriz Haupt/Friedrich Vogt/Carl von Kraus bearbeitet von Hugo Moser/Helmut Tervooren, 36. neugestaltete und erweiterte Aufl., Stuttgart 1977.

SM - Die Schweizer Minnesänger, nach der Ausg. von Karl Bartsch neu bearbeitet und hg. von Max Schiendorfer, Tübingen 1990.

W - Walther von der Vogelweide: Leich, Lieder, Sangsprüche, 14. völlig neubearbeitete Aufl. der Ausgabe von Karl Lachmann mit Beiträgen von Thomas Bein und Horst Brunner hrsg. von Christoph Cormeau, Berlin/New York 1996.

## Sekundärliteratur

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

Box, George E. P. Robustness in the Strategy of Scientific Model Building, in: Launer, Robert L./Wilkinson, Graham N. (Hrsg.): Robustness in Statistics, New York [u. a.] 1976, S. 201–236.

Braun, Manuel/Reiter, Nils: Sangsprüche auf/in Wörterwolken oder: Vorläufige Versuche zur Verbindung quantitativer und qualitativer Methoden bei der Erforschung mittelhochdeutscher Lyrik, in: Brunner, Horst/Löser, Freimut/Franzke, Janina (Hrsg.): Sangspruchdichtung zwischen Reinmar von Zweter, Oswald von Wolkenstein und Michel Beheim, Wiesbaden 2017 (Jahrbuch der Oswald von Wolkenstein-Gesellschaft 21), S. 5–20 (online).

Burrows, John: All the Way Through. Testing for Authorship in Different Frequency Strata, in: Literary and Linguistic Computing 22 (2007), S. 27–47 (online).

Ciula, Arianna/Eide, Øyvind/Marras, Cristina/Sahle, Patrick (Hrsg.): Models and Modelling between Digital and Humanities: A Multidisciplinary Perspective. Historical Social Research, Supplement 31 (2018) (online).

Craig, Hugh/Greatley-Hirsch, Brett: Style, Computers, and Early Modern Drama Beyond Authorship, Cambridge 2017.

Eder, Maciej/Kestemont, Mike/Rybicki, Jan: Stylometry with R: a suite of tools, in: Digital Humanities 2013: Conference Abstracts. University of Nebraska-Lincoln, 16. –19. July 2013, NE, S. 487–489 (online).

Escobar Varela, Miguel: Theatre as Data. Computational Journeys into Theater Research, Ann Arbor 2021 (online).

Fellows, Ian: wordcloud: Word Clouds. R package 2018 (online).

Firth, John R.: A synopsis of linguistic theory 1930–1955, in: Firth, John R. (Hrsg.): Studies in Linguistic Analysis. Special volume of the Philological Society, Oxford 1957, S. 1–32.

Flanders, Julia/Jannidis, Fotis (Hrsg.): The Shape of Data in Digital Humanities. Modeling Texts and Text-based Resources, London 2018.

Gius, Evelyn/Jacke, Janina: Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse, in: Zeitschrift für digitale Geisteswissenschaften 1 (2015) (online).

Grün, Bettina/Hornik, Kurt: topicmodels: An R Package for Fitting Topic Models 2011 (online).

Horstmann, Jan: Topic Modeling, in: *forTEXT. Literatur digital erforschen*, 2018 (online).

Hübner, Gert: Minnesang im 13. Jahrhundert. Eine Einführung, Tübingen 2008.

Hübner, Gert: Konzentration aufs Kerngeschäft. Späte Korpora der Manessischen Liederhandschrift und die Gattungsgeschichte des Minnesangs im 13. Jahrhundert, in: Köbele 2013a, S. 387–411.

Jannidis, Fotis/Flanders, Julia: A Gentle Introduction to Data Modeling, in: Flanders/Jannidis 2018, S. 26-96.

Jannidis, Fotis: Modeling in the Digital Humanities: a Research Program?, in: Ciula [u. a.] 2018, S. 96–100 (online).

Karsdorp, Folgert/Kestemont, Mike/Riddell, Allen: Humanities Data Analysis. Case Studies with Python, Princeton 2021.

Klein, Thomas/Wegera, Klaus-Peter/Dipper, Stefanie/Wich-Reif, Claudia (2016): Referenzkorpus Mittelhochdeutsch (1050–1350), Version 1.0 (online).

Köbele, Susanne (Hrsg., in Verbindung mit Eckart Conrad Lutz und Klaus Ridder): Transformationen der Lyrik im 13. Jahrhundert. Berlin 2013a (Wolfram-Studien 21).

Köbele, Susanne (2013b): Einleitung, in: Köbele 2013a, S. 9–17.

Kuhn, Hugo: Minnesangs Wende, Tübingen 1952.

McCarthy, Philip M.: An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity. Dissertation University of Memphis 2005.

McCarty, Willard: Modeling: A Study in Words and Meanings, in: Schreibman, Susan/Unsworth, John/Siemens, Ray (Hrsg.): A Companion to Digital Humanities, Oxford (2004) (online).

Moretti, Franco: Conjectures on World Literature, in: New Left Review 1 (2000), S. 54–68 (online).

Moretti, Franco: »Operationalizing«: or, the function of measurement in modern literary theory. Literary Lab Pamphlet 6 (2013) (online).

Mueller, Martin: Shakespeare His Contemporaries. Collaborative curation and exploration of Early Modern drama in a digital environment, in: Digital Humanities Quarterly 8 (2014) (online).

Perkhun, Rainer/Keibel, Holger/Kupietz, Marc: Ergänzungen zu Korpuslinguistik. 18. Juni 2012 (online).

Pierazzo, E[lena]: How Subjective Is Your Model?, in: Flanders/Jannidis 2018, S. 117–132.

Piper, Andrew: There Will Be Numbers, in: Cultural Analytics 1 (2016), S. 1–10 (online).

Piper, Andrew: Think Small: On Literary Modeling, in: PMLA 132 (2017), S. 651–658 (online).

Ramsay, Stephen: Reading Machines. Toward an Algorithmic Criticism, Champaign 2011 (online).

Rosen, Jeremy: Combining Close and Distant, or, the Utility of Genre Analysis: A Response to Matthew Wilkens's »Contemporary Fiction by the Numbers«, in: Post45 2011 (online).

Schmid, Helmut: Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts, in: Proceedings DATeCH, May 2019 (online).

Schnell, Rüdiger: Minnesang und Sangspruch im 13. Jahrhundert. Gattungsdifferenzen und Gattungsinterferenzen, in: Köbele 2013a, S. 287–347.

Schöch, Christof: Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, in: Digital Humanities Quarterly 11 (2017), H. 2 (online).

Schöch, Christof: Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie, in: Bernhart, Toni/Willand, Marcus/Richter, Sandra/Albrecht, Andrea: Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven, Berlin/Boston 2018, S. 77–94 (online).

Shen Y[an] S[hun], Lucas/Lesieur, David/Bedetti, Christophe: LexicalRichness: A small module to compute textual lexical richness 2022 (online).

So, Richard Jean: »All Models Are Wrong«, in: PMLA 132 (2017), S. 668–673 (online).

Spärck Jones, Karen: A Statistical Interpretation of Term Specificity and Its Application in Retrieval, in: Journal of Documentation 28 (1972), 11–21 (online).

Stachowiak, Herbert: Allgemeine Modelltheorie, Wien 1973.

Underwood, Ted: A Genealogy of Distant Reading, in: Digital Humanities Quarterly 11 (2017) (online).

Viehhauser, Gabriel: Digitale Gattungsgeschichten. Minnesang zwischen generischer Konstanz und Wende, in: Zeitschrift für digitale Geisteswissenschaften (2017) (online).

Viehhauser, Gabriel: Mittelalterliche Texte als Modellierungsaufgabe, in: Fischer, Martin (Hrsg.): Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen. Akten der Tagung Bamberg, 08.–10. November 2018, Bamberg 2020, S. 15–50 (online).

Wiedemann, Gregor/Niekler, Andreas: Hands-on: A five day text mining course for humanists and social scientists in R. Proceedings of the 1st Workshop on Teaching NLP for Digital Humanities (Teach4DH@GSCL 2017), Berlin 2017 (online).

Windelband, Wilhelm: Geschichte und Naturwissenschaft. Rede zum Antritt des Rectorats der Kaiser-Wilhelms-Universität Strassburg, geh. am 1. Mai 1894. Strassburg 1894. Sitzungsberichte der Heidelberger Akademie der Wissenschaften, Philosophisch-Historische Klasse. Jg. 1910, Abh. 14 (online).

## Online-Ressourcen

MHDBDB (Mittelhochdeutsche Begriffsdatenbank): http://mhdbdb.sbg.ac.at/.

Python-wordcloud: https://github.com/amueller/word_cloud.

RNNTagger (Recurrent Neural Network Tagger): https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/.

scipy.stats.linregress: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html.

TDS (Towards Data Science): https://towardsdatascience.com/how-to-make-word-clouds-in-python-that-dont-suck-86518cdcb61f.

## Anschrift des Autors:

Prof. Dr. Gabriel Viehhauser
Universität Stuttgart
Institut für Literaturwissenschaft
Herdweg 51
70174 Stuttgart
E-Mail: viehhauser@ilw.uni-stuttgart.de