Methoden der Digitalen Humanwissenschaften anhand eines Korpus zeitgenössischer Lyrik

Julian Scheuchenpflug, MatNr.: 2000910, Mail: julian.scheuchenpflug@stud-mail.uni-wuerzburg.de,
PO 2015
Modul: 04-DH-E2-152, Seminar: Forschungsmethoden, Dozent: Prof. Dr. Fotis Jannidis

Inhalt

Einleitung

Schreiben und Textproduktion sind zweifellos von einer breiten Palette unterschiedlicher Faktoren beeinflusst, die weit über die bloße Beherrschung der Sprache hinausgehen. Diese Faktoren sind vielschichtig und komplex und schließen soziokulturelle Variablen wie Herkunft, Erziehung, Enkulturation, Alter, Geschlecht und viele andere ein. Die Formulierung dieser Aussage mag zunächst absichtlich diffus erscheinen, doch genau hierin liegt der Ausgangspunkt für eine tiefergehende Analyse, die nur durch sorgfältige Untersuchungen und wissenschaftlich-methodische Präzision in kleinen Schritten bestätigt oder widerlegt werden kann.

Das Hauptziel dieser Forschungsarbeit besteht darin, Theorien im Kontext von Schreiben und Textproduktion zu dekonstruieren und sie mithilfe präziser wissenschaftlicher Methoden, einschließlich statistischer Tests, genauer zu untersuchen. Diese Methodik erfordert nicht nur ein Verständnis der vorhandenen Literatur und Forschung, sondern auch die Fähigkeit, diese Erkenntnisse in strukturierte Hypothesen umzuwandeln. Der Forschungsprozess beginnt damit, bestehende Theorien und wissenschaftliche Erkenntnisse in diesem Bereich zu analysieren, und aus diesen dann spezifische Theorien abzuleiten, die als Grundlage für die weiteren Untersuchungen dienen. Diese Theorien werden in präzise Hypothesen umgewandelt, die klar formuliert und testbar sind. Diese Hypothesen werden dann anhand eines sorgfältig ausgewählten Korpus getestet, der für die Forschungsfrage von Relevanz ist. Die Datenerhebung und -analyse erfolgen mit wissenschaftlicher Präzision, um valide Ergebnisse zu gewährleisten. Diese Arbeit trägt dazu bei, das Verständnis darüber zu vertiefen, wie sich wissenschaftliche Methodik gestaltet und beschäftigt sich mit dem Prozess wissenschaftlicher Erkenntnisfindung. Darüber hinaus fördert sie das Verständnis für die Bedeutung soziokultureller Einflüsse auf sprachliche Prozesse. Insgesamt könnte diese Forschungsarbeit dazu beitragen, die Komplexität von Schreiben und Textproduktion in der vorliegenden Stichprobe und den hierbei verwendeten Metriken aufzuschlüsseln.

Fragestellung, Theorien und Hypothesen

In dem dieser Arbeit vorangegangenen Gruppenprojekt wurde bereits eine grundlegende Analyse der Struktur sowie formellen und inhaltlichen Metriken eines Teils dieser Gedichte durchgeführt. Die Ergebnisse dieser Untersuchung sollen hier noch einmal kurz zusammengefasst werden. Es ist jedoch wichtig zu betonen, dass zum Zeitpunkt der Projektbearbeitung keine Informationen zur zeitlichen Einordnung der Gedichte verfügbar waren, wie beispielsweise das Geburtsjahr der Autor*innen oder das Veröffentlichungsjahr der Gedichte. Daher wird diese Arbeit nun einen tieferen Blick auf das Thema Zeit werfen und untersuchen, wie sich die bereits

erarbeiteten Ergebnisse mit der bislang unbehandelten Variablen Zeit verknüpfen lassen. Die Grundtheorie, die zu Beginn der Arbeit erwähnt wurde, scheint im aktuellen wissenschaftlichen Diskurs durchaus relevant zu sein. Soziokulturelle und insbesondere erziehungsspezifische Faktoren, die das Schreiben und die Textproduktion beeinflussen können, sind eng mit der jeweiligen Zeitperiode verbunden. In diesem Zusammenhang wird besonders auf die Unterschiede in der Erziehung von Jungen und Mädchen bzw. Männern und Frauen eingegangen, ohne dabei extreme Präzision anzustreben. Dies kann allein anhand der rechtlichen Entwicklungen im 20. Jahrhundert verdeutlicht werden. Die Emanzipation der Frau im 20. Jahrhundert sowie die kontinuierliche Entwicklung des deutschen Bildungssystems sind Themen, die umfangreiche Studien erfordern und in dieser Arbeit nicht im Fokus stehen. Dennoch soll dieser geschichtliche Hintergrund als Aufhänger für die hier angestrebte Hypothesenbildung dienen. Die vorliegende Untersuchung wird sich stattdessen darauf konzentrieren, wie sich diese soziokulturellen und erziehungsspezifischen Faktoren im Kontext der Zeitveränderungen manifestieren. Wie haben sich Schreibstile in Gedichten im Laufe der Jahrzehnte geändert? Gibt es Trends oder Muster, die auf Veränderungen in der Gesellschaft, der Bildung oder der Rolle von Frauen und Männern hinweisen? Diese Fragen werden im Zentrum der weiteren Forschungsarbeit stehen und sollen dazu beitragen, das Verständnis für die Wechselwirkungen zwischen Literatur, Zeit und Gesellschaft zu vertiefen. Die grundlegende Fragestellung lautet also, ob sich das Schreiben von Männern vom Schreiben von Frauen unterscheidet. In dieser Arbeit werden mehrere Hypothesen formuliert, um Geschlechterunterschiede in der Gedichtproduktion zu untersuchen. Die Hypothesen basieren auf der Annahme, dass Frauen und Männer unterschiedliche Schreibstile und -muster aufweisen können, die auf soziokulturelle Faktoren wie Erziehung und Bildung zurückführbar sein könnten. Im Folgenden werden die Hypothesen H1, H2 und H3 im Detail dargestellt:

H1: Frauen schreiben längere Gedichte als Männer.

Die erste Hypothese basiert auf der Vermutung, dass es Unterschiede in der Länge von Gedichten gibt, die von Frauen und Männern verfasst werden. Frauen könnten tendenziell längere Gedichte schreiben als Männer. Die Gedichtlänge wird in dieser Arbeit als einfachste Metrik verwendet, um diese Hypothese zu überprüfen. Darüber hinaus werden auch die Verslängen pro Gedicht, sowie über die gesamten weiblichen bzw. männlichen Korpora mit einbezogen.

H2: Frauen nutzen mehr beschreibende Wörter in ihren Gedichten als Männer.

Die zweite Hypothese zielt darauf ab, festzustellen, ob Frauen in ihren Gedichten häufiger beschreibende Wörter verwenden als Männer. Beschreibende Wörter umfassen hier Adjektive. Die einfache Metrik also, die zur Überprüfung dieser Hypothese verwendet wird, ist der relative

Adjektivanteil, der das Verhältnis der beschreibenden Wörter zu anderen Worten im Gedicht misst.

H3: Männer nutzen mehr benennende Wörter in ihren Gedichten als Frauen.

Als Gegenüberstellung zur zweiten Hypothese wird in der dritten Hypothese angenommen, dass Männer in ihren Gedichten tendenziell mehr benennende Wörter verwenden. Benennende Wörter umfassen hier Substantive und dienen dazu, konkret benennbare Objekte oder Konzepte im Gedicht zu präsentieren. Die einfache Metrik, um diese Hypothese zu überprüfen, ist der Substantivanteil, der das Verhältnis von benennenden Wörtern zu anderen Wortarten im Gedicht misst.

H4: Die Hypothesen H2 und H3 korrelieren mit dem Geburtsdatum der Autor*innen.

Die vierte Hypothese untersucht, ob die in den Hypothesen H2 und H3 festgestellten Geschlechterunterschiede in der Verwendung von beschreibenden und benennenden Wörtern mit den Geburtsdaten der Autor*innen korrelieren. Es wird angenommen, dass Bildung und Erziehung einen Einfluss auf den Schreibstil haben können und somit auch auf die Verwendung von bestimmten Wortarten in Gedichten. Das Geburtsdatum der Autor*innen wird als Indikator für die zeitliche Dimension dieser Hypothese verwendet.

In der vorliegenden Arbeit werden statistische Analysen durchgeführt, um diese Hypothesen zu testen und festzustellen, ob signifikante Geschlechtsunterschiede in der Gedichtproduktion und im Schreibstil existieren.

Das Korpus

Das vorliegende Korpus stellt eine Sammlung von 2375 zeitgenössischen Gedichten deutscher Sprache dar, die von verschiedenen Autorinnen und Autoren mit Geburtsjahren im 19., 20. und 21. Jahrhundert verfasst wurden. Hierbei stammen 938 Gedichte von Autorinnen und 1437 von Autoren. Diese Gedichte wurden von der öffentlich zugänglichen Website lyrikline.org extrahiert. Neben den eigentlichen Gedichttexten bietet die Website auch wertvolle biographische Informationen über die Autor*innen. Es ist interessant festzustellen, dass nur wenige Autor*innen auf dieser Plattform mehr als fünf Gedichte veröffentlicht haben, was die Vielfalt der Schreibenden und ihrer Werke unterstreicht.

Datenakquise

Die Datenvorbereitung für dieses Korpus war keine einfache Aufgabe und erforderte den Einsatz von Web-Scraping-Tools wie *BeautifulSoup* und *Selenium*, um die Gedichte und zugehörigen Informationen von der Website zu extrahieren. Die anschließende Tokenisierung der Gedichte sowie das POS-Tagging wurden mithilfe von *Spacy* 3.5.3 und dem Modell *de-core-news-*

md 3.5.0 durchgeführt. Allerdings gestaltete sich die Ermittlung der Veröffentlichungsdaten der Gedichte als problematisch. Oft waren diese Daten unvollständig oder lediglich mit dem Datum der Audioproduktion auf lyrikline.org versehen. Hier konnte keine zufriedenstellende Lösung gefunden werden, um die Veröffentlichungsdaten präzise und ohne erheblichen Aufwand zu extrahieren, ohne die Gedichte einzeln auf Veröffentlichungsdaten in verschiedensten Publikationen zu prüfen. Aus diesem Grund wurde das Geburtsjahr der Autor*innen als Ersatz für die Veröffentlichungsdaten verwendet. Diese Herangehensweise ist zweifellos fehleranfällig und kann keine genauen Rückschlüsse auf stilistische Veränderungen in den Gedichten im Laufe der Jahre und Jahrzehnte bieten. Sie dient jedoch als Ausgangspunkt für eine Analyse im Hinblick auf die pädagogische und akademische Erziehung der einzelnen Autor*innen. Auch nach dem vorangegangenen Scraping wurden nur Gedichtdateien verwendet, die sowohl ein mit Text gefülltes Gedicht als auch ein gültiges Geburtsjahr der Autor*innen enthalten.

Die vorliegende Datenbank mit den zeitgenössischen Gedichten und den zugehörigen Informationen zu den Autor*innen ist eine wertvolle Ressource für die Literaturforschung und bietet die Möglichkeit, tiefgehende Analysen im Bereich der Literaturwissenschaft und der Poesie durchzuführen. Sie ermöglicht Einblicke in die Vielfalt literarischer Ausdrucksformen und die Art und Weise, wie verschiedene Autor*innen verschiedene Themen in ihren Werken interpretieren. Trotz der Herausforderungen bei der Datenvorbereitung ist dieses Korpus ein wichtiges Instrument zur Vertiefung des Verständnisses für die Beziehung zwischen Literatur und Zeit.

Begriffsbildung und Definitionen

Zunächst soll sich hier mit der Begriffsbildung im Verlauf der Forschung auseinandergesetzt werden. Glücklicherweise sind die Begriffe im Kontext der rein formalen Analyse von Gedichten relativ einfach und klar definiert. Zunächst zum Gedicht selbst: Gedichte sind eine *„dichterische, meist gereimte Kunstform, die durch ihren Rhythmus und eine Gliederung in Verse und Strophen bestimmt ist und dem Leser oder Hörer vor allem eine Stimmung, ein seelisches Erlebnis vermittelt“*[1]. Gerade in Bezug auf zeitgenössische, modernere Gedichte und experimentelle Lyrik, von denen im behandelten Korpus bei genauer Betrachtung recht viele zu finden sind, ist die fehlende Trennschärfe in der obenstehenden Definition von Vorteil. Viele der vorliegenden Gedichte sind nicht gereimt und verfügen über keinen festen Rhythmus im

[1] https://www.dwds.de/wb/Gedicht [aufgerufen am 22.09.2023]

klassischen Sinne von Jambus, Trochäus, etc. Zum Beispiel in dem Gedicht *** [أرض شفافة — Maervent Oiosis] von Daniel Falb[2]:

„***

أرض شفافة — Maervent Oiosis.

Die geothermale Quelle von
أرض

Hallo, Lsx,

Hier meine Euro-Imagination

von unaufgegessenem lachs,
Fruchtsaft
die rechnenden Granulate, — Fortec, Landec
Ultra Cedorum.. —,
— Sahaelgürtel bei 27° nördl.,
EExai" III, s.S. 34 —

— es könnte so gewesen sein —
Enjoy!
Hallo, Tüten aus recyceltem Papier mit Fortec stehen auf Ihren Metallregalen,
im Keller

Mit 1 kg davon liegt man bei um die 5.4 × 1050 OPS
(Seth Lloyd, UNLESERLICHES GEKRITZEL´)
Umgekehrt, bei 8.99 × 1016 J pro kg und insgesamt 2,6 kg kommt man auf 5,58 × 1016 cal
in allen Tüten zusammen, in denen das Klimamodell des dunklen Kellerraums läuft
inklusive der in FortecTM selbst entstehenden Wärme, über Verwirbelungen
darin, wenn Sie Lsx ,mit abgewetztem Jackett, braunen hornigen Zehennägeln...
auf dem Lehmboden zwischen den Regalen umherlaufen und einige Tüten umfüllen,
bis hin zur feuchten, kriechenden Fahne über dem Quell von أرض
mit installierter geotherm. Leistung P
=====23,36 × 1016- Ein
enim aiont landec!! —

für gutes Wachstum der

Schule für Erde, Energie und Umweltwissenschaften!

— Anker-Registrierkasse, —
der alte, alte *..

. ,
(Kinderkaufmannsladen) ◊
Kalorimeter
C-84.
أرض شفافة — Maervent Oiosis animertoutlanding,
Bitte, Die E.c.d.e.

des Denkens.“

[2] Falb, Daniel: *** [أرض شفافة — Maervent Oiosis]. Berlin: kookbooks 2015. Online unter https://www.lyrikline.org/en/poems/maervent-oiosis-11952 [aufgerufen am 22.09.2023]. Die auf der Website dargestellte Formatierung, Rechtschreibung und Nutzung von Sonderzeichen ist hier weitestgehend erhalten geblieben.

Hier wird die Definition des DWDS also maßgeblich beschnitten, und könnte in der weiteren Arbeit wohl so formuliert werden: *„dichterische [..] Kunstform, die durch [..] eine Gliederung in Verse [..] bestimmt ist und dem Leser oder Hörer vor allem eine Stimmung, ein seelisches Erlebnis vermittelt“*. Weiterhin wichtig für die durchgeführten Analysen am Korpus ist die Definition von „Vers“: *„metrisch, rhythmisch gestaltetes, oft gereimtes sprachliches Gebilde in gebundener Rede, das meist eine Zeile in einem Gedicht, einer Strophe, in einem Drama oder Epos bildet“*[3]. Hierbei wird schon bei der Datenextraktion ein bedeutender Einschnitt gemacht: die Unterteilung in Verse findet hier durch die auf der Website gesetzten
-Elemente statt. Folgendes Beispiel[4] zeigt unter anderem die Versaufteilung wie von der Website extrahiert[5]:


```
"line.1": {
            "text": "wir wissen schon alles."
        },
"line.2": {
            "text": "die welt besteht."
        },
"line.3": {
            "text": "die wildnisse breiten sich aus."
        },
"line.4": {
            "text": "der vater zeugt den sohn, das bedeutet: er krallt sich ins fleisch ei-
ner oiden in einer weise, dass er sie auch dann noch wird zreissen, wenn er nicht mehr da-
neben ist, und dass sie wird wissen, es war er und kein anderer, der sie zrissen hat. und
dann ist da der sohn in einer der wildnisse und frißt, was von der mutter übrig ist, und
schaut prüfend den himmel an, und schaut prüfend die äste der niedrigen bäume an, und sieht
in der ferne, denn so funktioniert dieses gedicht, den horizont, der der vater ist, und
weil er als das zreissen der welt in der welt ist, das der vater et cetera, so schickt er
sich an, den vater zu zreissen."
        },
"line.5": {
            "text": "sie kämpfen."
        },
```


Diese wurde unter der Annahme vorgenommen, dass die Aufteilung auf der Website den Wünschen und Vorstellungen der Autor*innen entspricht und somit unabhängig der nutzersichtbaren Formatierung eine Verseinteilung repräsentiert.

[3] https://www.dwds.de/wb/Vers [aufgerufen am 22.09.2023]
[4] Schmitzer, Stefan: #6 (zeus kronion). O.J. Online unter: https://www.lyrikline.org/en/poems/6-zeus-kronion-16611 [aufgerufen am 22.09.2023]
[5] Die Repräsentation im Korpus.

Analyse der Daten

In der statistischen Analyse wurden verschiedene Merkmale in Bezug auf männliche und weibliche Autoren untersucht, um die vorher genannten Hypothesen zu testen. Dabei wurde zunächst mithilfe des Shapiro-Wilk-Tests überprüft, ob die Daten normalverteilt sind. Wenn die Daten normalverteilt sind, wird in der Regel ein t-Test angewendet, andernfalls der Mann-Whitney-U-Test. Zunächst soll die Hypothese H1 getestet werden:

Gedichtlängen (Tokens pro Gedicht):

- männlich: Durchschnittliche Gedichtlänge: 160.320, Standardabweichung: 153.715, Stichprobengröße: 1459
- weiblich: Durchschnittliche Gedichtlänge: 137.258, Standardabweichung: 134.297, Stichprobengröße: 950

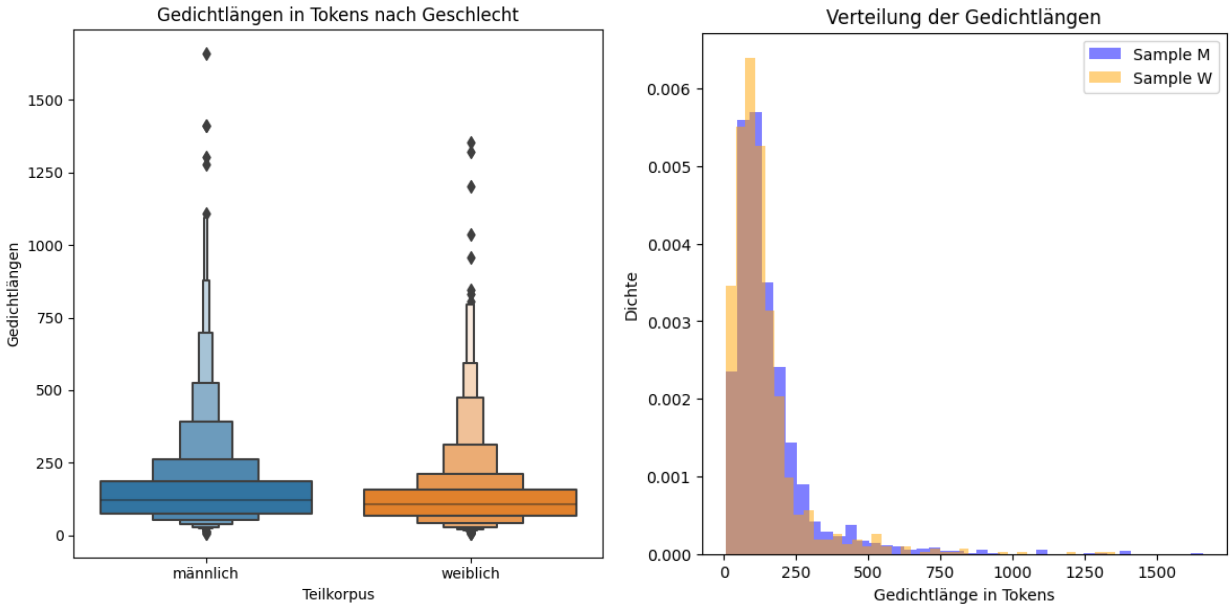


Abbildung 1: links: Gedichtlängen nach Tokens, rechts: Verteilung der Gedichtlängen im Korpus

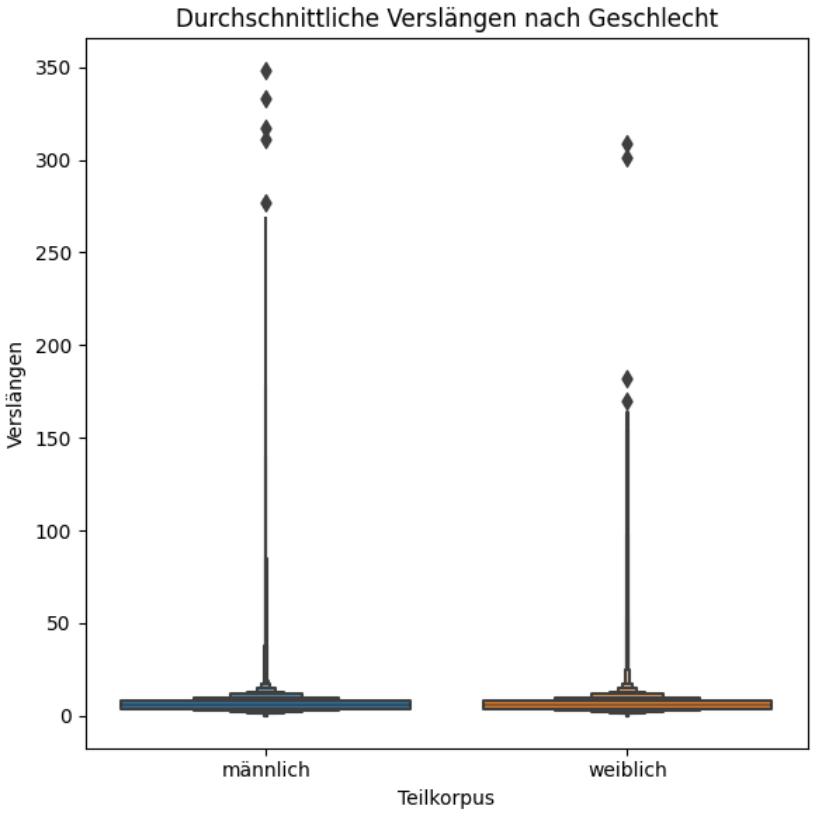
Da die Gedichtlängen nicht normalverteilt sind, wurde erneut der Mann-Whitney-U-Test angewendet. Das Ergebnis zeigte einen signifikanten Unterschied zwischen den Subsamples ($U = 789737.5$, $N(m) = 1437$, $N(w) = 938$, $p < 0.001$) was darauf hinweist, dass es einen statistisch signifikanten Unterschied in den Gedichtlängen zwischen männlichen und weiblichen Autoren gibt.

Tokens pro Vers im Gesamtkorpus:

- männlich: Durchschnittliche Tokens pro Vers: 6.516, Standardabweichung: 6.502, Stichprobengröße: 35884

- weiblich: Durchschnittliche Tokens pro Vers: 6.336, Standardabweichung: 6.470, Stichprobengröße: 21004

Da die Tokens pro Vers nicht normalverteilt sind, wurde der Mann-Whitney-U-Test angewendet. Das Ergebnis zeigte einen signifikanten Unterschied zwischen den Subsamples (U = 394549287.0, N(m) = 35310, N(w) = 20369, $p < 0.001$), was darauf hinweist, dass es einen statistisch signifikanten Unterschied in den Tokenanzahlen pro Vers zwischen männlichen und weiblichen Autoren gibt. Die tatsächliche, absolute Abweichung der Verslänge ist aber sehr gering.



- männlich: Durchschnittliche Verslänge: 8.093, Standardabweichung: 15.847
- weiblich: Durchschnittliche Verslänge: 9.221, Standardabweichung: 21.366

Abbildung 2: Durchschnittliche Verslängen nach Gedicht

Auch hier waren die Daten nicht normalverteilt, der Mann-Whitney-U-Test ergab keinen signifikanten Unterschied zwischen den Subsamples (U = 724864.5, N(m) = 1459, N(w) = 970, p = 0.308).

Zusammenfassend zeigen die Analysen die erste Hypothese betreffend, dass es signifikante Unterschiede in den Tokenfrequenzen pro Vers und den Gedichtlängen zwischen männlichen und weiblichen Autoren gibt, während es keinen signifikanten Unterschied in den Verslängen pro Gedicht gibt. Interessant ist hierbei, dass männliche Autoren im vorliegenden Korpus längere Gedichte schreiben, sowie generell auf das Gesamtkorpus gesehen längere Verse, die Verslänge

pro Gedicht aber leicht kürzer und auch weniger weit verteilt ist als bei den weiblichen Autorinnen.

Die Hypothese H2 wurde mit der durchschnittlichen Frequenz von Adjektiven für den gesamten Korpus bei männlichen und weiblichen Autoren getestet. Bei männlichen Autoren betrug die durchschnittliche Adjektivfrequenz 0.0372, mit einer Standardabweichung von 0.028. Im Gegensatz dazu lag die durchschnittliche Adjektivfrequenz bei weiblichen Autoren bei 0.0427, wobei die Standardabweichung 0.032 betrug. Für den Gesamtkorpus wurde eine durchschnittliche Adjektivfrequenz von 0.0393 ermittelt.

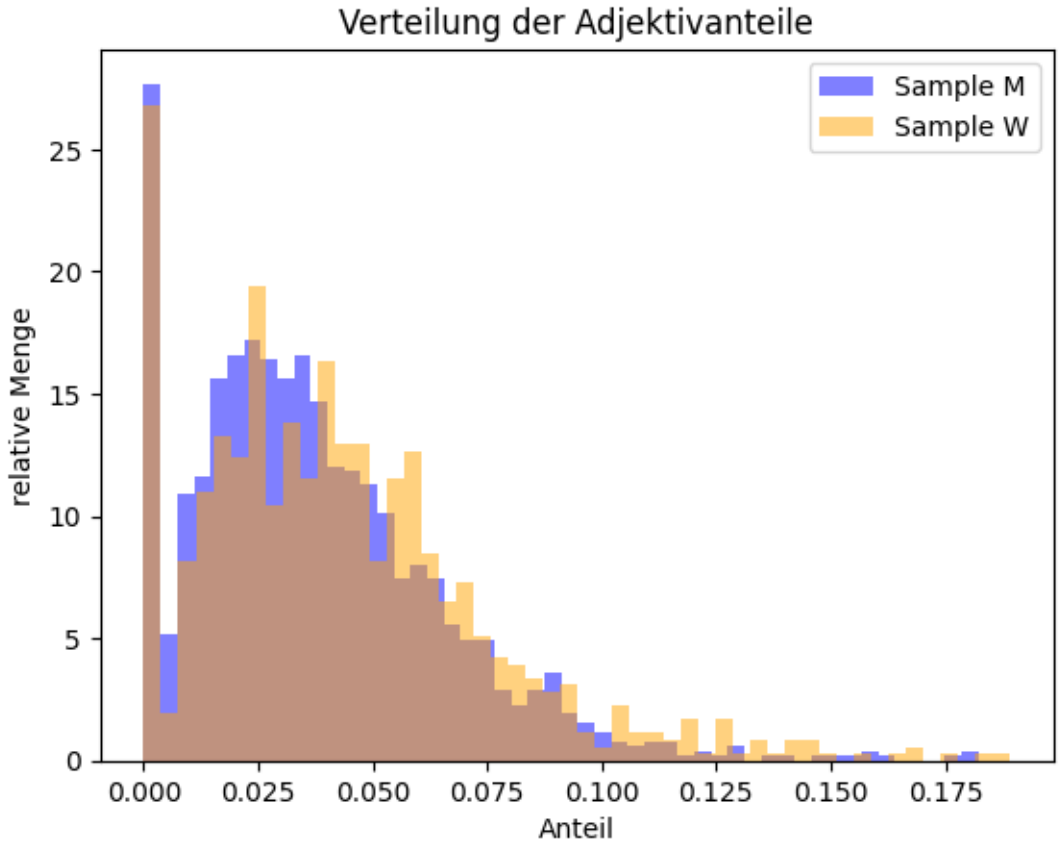


Abbildung 3: Verteilung der Adjektivanteile

Um festzustellen, ob es einen signifikanten Unterschied in den Adjektivfrequenzen zwischen männlichen und weiblichen Autoren im gesamten Korpus gibt, wurde der Mann-Whitney-U-Test durchgeführt. Hier zeigt sich ein signifikanter Unterschied in den Adjektivfrequenzen zwischen den beiden Geschlechtern im gesamten Korpus ($U = 608502$, $N(m) = 1437$, $N(w) = 938$, $p < 0.001$). Die Hypothese H2 kann also somit vorläufig akzeptiert werden. Im Gegensatz zur nur minimalen Abweichung der absoluten Werte bei den Tokens pro Vers kann bei dieser prozentualen Angabe der Adjektivfrequenz auch von einem „spürbaren“ Unterschied gesprochen werden, der berechnete Unterschied von 0.55 Prozentpunkten bzw. 14,78% scheint doch ausschlaggebend.

Zur Prüfung von H3 wurde die durchschnittliche Frequenz von Substantiven für den gesamten Korpus sowohl bei männlichen als auch bei weiblichen Autoren untersucht. Bei männlichen Autoren betrug die durchschnittliche Substantivfrequenz 0.201, mit einer Standardabweichung von 0.064. Im Gegensatz dazu lag die durchschnittliche Substantivfrequenz bei weiblichen Autoren bei 0.206, mit einer Standardabweichung von 0.062. Für den Gesamtkorpus, der beide Gruppen umfasst, ergab sich eine durchschnittliche Substantivfrequenz von 0.203.

Um festzustellen, ob es einen signifikanten Unterschied in den Substantivfrequenzen zwischen männlichen und weiblichen Autoren im gesamten Korpus gibt, wurde der Mann-Whitney-U-Test durchgeführt. Hierbei sollte beachtet werden, dass die Daten im unten stehenden Graphen zwar annähernd normalverteilt anmuten und somit die Nutzung des t-Tests rechtfertigen könnten, diese Annahme aber durch einen Shapiro-Wilk-Test abgelehnt werden konnte.

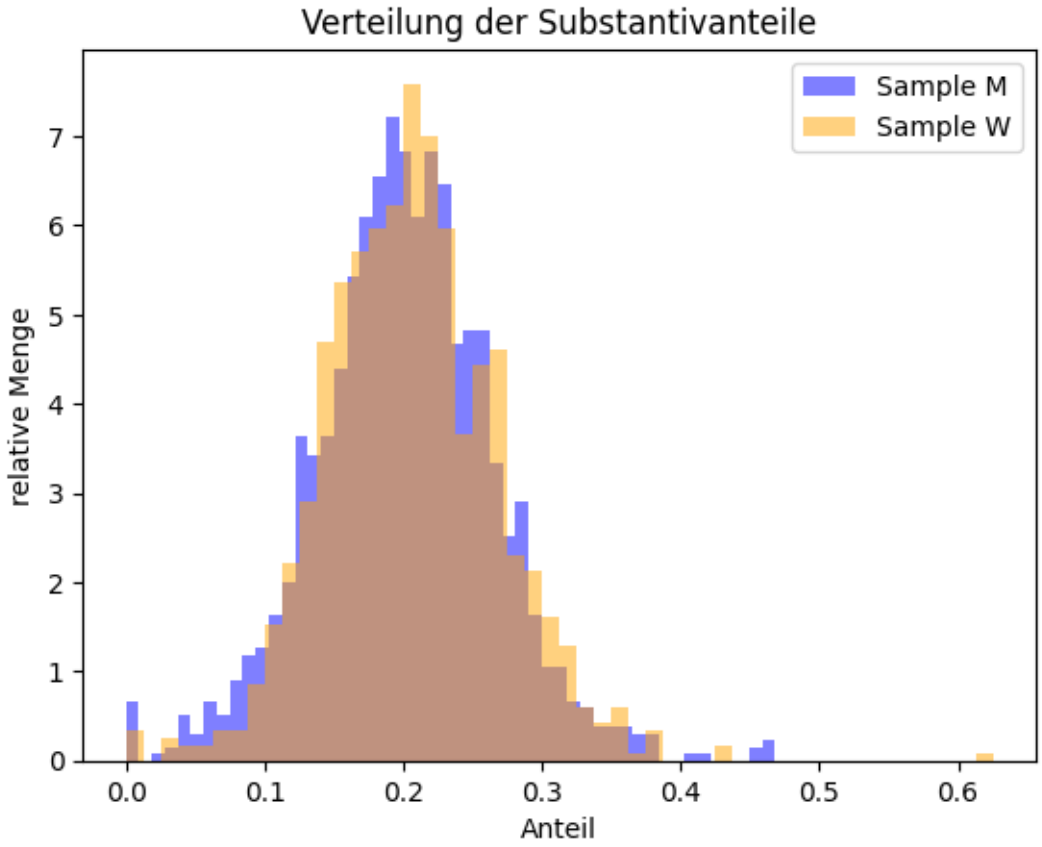


Abbildung 4: Verteilung der Subjektivanteile

Die Teststatistik ergab keinen signifikanten Unterschied in den Substantivfrequenzen zwischen den beiden Geschlechtern im gesamten Korpus (U = 648942.0, N(m) = 1437, N(w) = 938, p = 0.126). Die Hypothese H3 ist somit vorläufig abgelehnt.

Die letzte hier behandelte Hypothese H4 untersucht die Unterschiede in Adjektiv- sowie Substantivfrequenzen im Bezug auf das Geburtsjahr der Autor*innen. Zu beachten sei hierbei, dass für jedes Subsample an Jahrzehnten der Shapiro-Wilk-Test zur Prüfung auf Normalverteilung

durchgeführt wurde, weswegen sich die ausgewählten Tests für einzelne Jahrzehnte unterscheiden können.

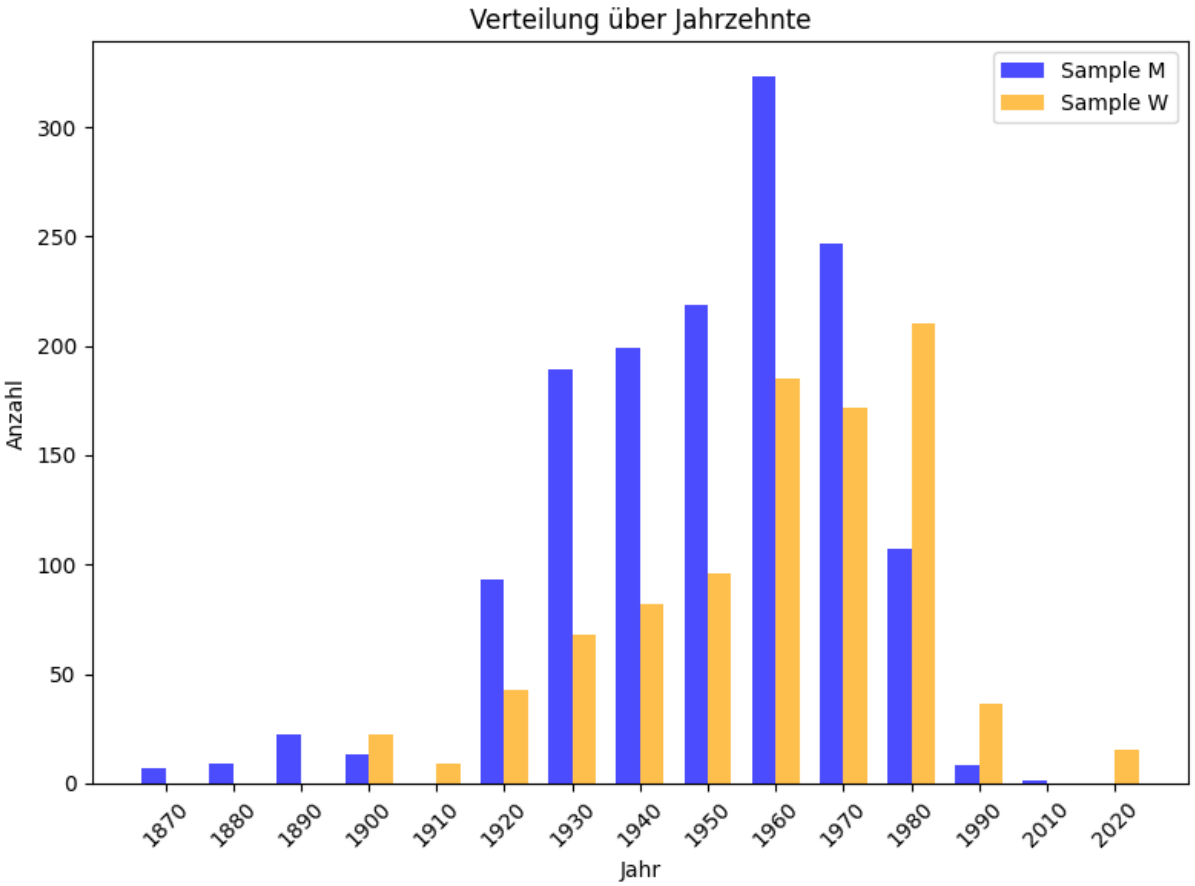


Abbildung 5: Verteilung der Gedichtanzahlen über die Jahrzehnte der Geburtsdaten

Für bestimmte Jahrzehnte wurden unterschiedliche Ergebnisse erzielt. In den 1920er Jahren gab es einen signifikanten Unterschied in den Substantivfrequenzen zwischen männlichen und weiblichen Autoren ($U = 1095$, $N(m) = 93$, $N(w) = 43$, $p < 0.001$). Für die 1940er, 1950er, 1960er, 1970er, 1980er und 1990er Jahre ergaben die jeweils gewählten Tests keinen signifikanten Unterschied. Es ist außerdem zu beachten, dass die 1870er, 1880er, 1890er, sowie 2010er und 2020er keine Tests durchgeführt werden konnten, da es weniger als 3 Datenpunkte für mindestens eine der Geschlechtergruppen gab.

Die Adjektivfrequenzen betreffend gab es in den 1920er Jahren einen signifikanten Unterschied zwischen männlichen und weiblichen Autoren gab ($U = 1521$, $p = 0.025$). Ähnliche Ergebnisse wurden für die 1940er ($U = 6375.5$, $N(m) = 199$, $N(w) = 82$, $p = 0.004$) und 1950er ($U = 8310.5$, $N(m) = 219$, $N(w) = 96$, $p = 0.003$) erzielt, wobei in beiden Fällen signifikante Unterschiede festgestellt wurden. Für andere Jahrzehnte ergaben die durchgeführten Tests keine signifikanten Unterschiede in den Adjektivfrequenzen zwischen den Geschlechtern.

Während diese Analyse zeigt, dass es in der Verwendung von Adjektiven zwischen männlichen und weiblichen Autoren im Korpus mehrere signifikante Unterschiede gibt, ergibt die Auswertung der Substantivanteile nur in den 1920er Jahren einen signifikanten Unterschied.

Diskussion und Fazit

Die Analyse zeigt deutliche Unterschiede in der Länge der Gedichte zwischen männlichen und weiblichen Autoren. Insbesondere bei den Adjektivfrequenzen wurden auffällige Ergebnisse festgestellt, die nicht nur statistisch signifikant sind. Dies weist darauf hin, dass die Häufigkeit von Adjektiven ein markantes Merkmal ist, das die Schreibstile der beiden Geschlechter voneinander unterscheidet. Es ist jedoch wichtig anzumerken, dass einige andere Variablen, die in dieser Untersuchung berücksichtigt wurden, keine signifikanten Unterschiede zwischen männlichen und weiblichen Autoren aufweisen. Dies kann darauf hinweisen, dass nicht alle sprachlichen Merkmale gleichermaßen von Geschlecht beeinflusst werden. Es gibt jedoch einige methodische Herausforderungen, die bei dieser Analyse berücksichtigt werden müssen. Erstens sind die Daten auf die verfügbaren Gedichte auf der Website beschränkt, und es ist unklar, wie diese Gedichte beschafft, hochgeladen und ausgewählt wurden. Dies könnte potenziell die Repräsentativität der Stichprobe beeinflussen. Zweitens besteht ein leichtes Ungleichgewicht in der Anzahl der Gedichte zwischen männlichen und weiblichen Autoren, was zu statistischen Herausforderungen führen kann. Drittens wurde ein relativ einfacher Tokenisierungs- und POS-Tagging-Prozess verwendet. Spacy bietet hier zwar ein breit anerkanntes Modell, trotzdem können Fehler in der Datenverarbeitung nicht ausgeschlossen werden, da Gedichte und besonders die vorliegenden zeitgenössischen Gedichte Merkmale aufweisen können, die nicht unbedingt den bekannten Regeln natürlicher Sprache unterliegen. Dies wirft auch die Frage auf, ob das sprachliche Modell von Spacy gegenüber Dialekten, nicht-deutschen Wörtern und anderen sprachlichen Nuancen stabil ist. Abschließend sei auch anzumerken, dass die Verwendung des Geburtsdatums der Autor*innen im Gegensatz zum tatsächlichen Veröffentlichungsdatum der Gedichte nicht ganz stabil den genannten TheorTrotz dieser Herausforderungen bieten die Ergebnisse dieser Analyse Einblicke in die Unterschiede und Gemeinsamkeiten in den Schreibstilen männlicher und weiblicher Autoren und unterstreichen die Komplexität der Sprachverwendung in literarischen Werken.

Quellen

https://www.dwds.de/wb/Gedicht [aufgerufen am 22.09.2023]

https://www.dwds.de/wb/Vers [aufgerufen am 22.09.2023]

Repository

https://github.com/niceshice/FM_SS2023_LL_Gedichte

Ich versichere, dass ich die Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sämtliche wörtlichen oder sinngemäßen Übernahmen und Zitate sind kenntlich gemacht und nachgewiesen.

Ferner versichere ich, dass das Thema dieser Arbeit nicht identisch ist mit dem Thema einer von mir bereits für eine andere Prüfung eingereichten Arbeit.

Ich erkläre weiterhin, dass ich die Arbeit nicht bereits an einer anderen Hochschule als Prüfungsleistung eingereicht habe.

Ravensburg, den 26.09.2023 _____

Datum, Unterschrift