

Beitrag aus:  
Zeitschrift für digitale Geisteswissenschaften

Titel:  
Gute Wörter, schwaches Gattungssignal. Differenzen zwischen Roman-Subgenres und Dramen mit Delta und signifikantem Wortschatz aufspüren

Autor\*in:  
Friedrich Michael Dimpel

Kontakt: [mail@dimpel.de](mailto:mail@dimpel.de)  
Institution: Friedrich-Alexander-Universität Erlangen-Nürnberg  
GND: [1111656460](#) ORCID: [0000-0003-4833-4897](#)


---

DOI des Artikels:  
[10.17175/2022\\_009\\_v2](https://doi.org/10.17175/2022_009_v2)

Nachweis im OPAC der Herzog August Bibliothek:  
[1866422553](#)

Erstveröffentlichung:  
17.11.2022

Version 2.0:  
14.11.2023

Lizenz:  
Sofern nicht anders angegeben 

Medienlizenzen:  
Medienrechte liegen bei den Autor\*innen

Letzte Überprüfung aller Verweise:  
30.10.2023

Format:  
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:  
[Gattungstheorie](#) | [Literaturgattung](#) | [Literaturwissenschaft](#) | [Statistische Stilistik](#) |

Empfohlene Zitierweise:  
Friedrich Michael Dimpel: Gute Wörter, schwaches Gattungssignal. Differenzen zwischen Roman-Subgenres und Dramen mit Delta und signifikantem Wortschatz aufspüren. In: Zeitschrift für digitale Geisteswissenschaften 7 (2022). 17.11.2022. Version 2.0 vom 14.11.2023. HTML / XML / PDF. DOI: [10.17175/2022\\_009\\_v2](https://doi.org/10.17175/2022_009_v2).

Änderungen in Version 2.0 (14.11.2023):  
Korrekturen entlang der Monita in den Gutachten: Sprachliche Verbesserungen; Ergänzungen in Kapitel 1 und 2 sowie in der Bibliografie; Ergänzung der Tabellenbeschriftungen, Aufschlüsselung von Abkürzungen

Friedrich Michael Dimpel

# Gute Wörter, schwaches Gattungssignal. Differenzen zwischen Roman-Subgenres und Dramen mit Delta und signifikantem Wortschatz aufspüren

---

## Abstracts

Untersucht wird, inwieweit sich die automatische Erkennung von Genres bzw. Subgenres mit Burrows' Delta durch signifikantes Vokabular (»Gute Wörter«) und Z-Wert-Begrenzung verbessern lässt. Auf einem Teilkorpus werden zu den Genres Abenteuerroman, Bildungsroman, Gesellschaftsroman, Komödie und Tragödie die »Guten Wörter« ermittelt; auf einem zweiten Teilkorpus wird evaluiert. Bei allen fünf Textsorten steigen die F1-Werte durch diese Optimierungsmaßnahmen, beispielsweise von 0,65 auf 0,77. Bei Abenteuerroman, Bildungsroman und Komödie steigen die F1-Werte beispielsweise von 0,79 auf 0,91. Die Klassifikation von Abenteuerroman versus Drama und von Komödie versus Abenteuer- und Bildungsroman gelingt fehlerfrei (ARI=1). Während das Gute-Wörter-Verfahren den Recall steigert, begrenzt die Z-Wert-Begrenzung die False-Positives.

It is investigated to what extent the automatic recognition of genres or subgenres by means of Burrows' Delta can be improved by significant vocabulary (»good words«) and Z-value limitation. On one subcorpus, »good words« are determined on the genres adventure novel, Bildungsromans, social novel, comedy, and tragedy; on a second subcorpus, they are evaluated. For all five text types, the F1 values increase due to these optimization measures, for example from 0.65 to 0.77. For adventure novel, Bildungsroman and comedy, the F1 values increase, for example, from 0.79 to 0.91. The classification of adventure novel versus drama and of comedy versus adventure and Bildungsroman succeeds without errors (ARI=1). While the »good word procedure« increases recall, the Z-score limitation limits false positives.

## 1. Gattung und Stilometrie

Während die Autorschaftserkennung auf digitalem Weg gut erforscht ist und sehr gute Erkennungsquoten vorweisen kann,<sup>1</sup> ist die digitale Erkennung von literarischen Gattungen deutlich anspruchsvoller.<sup>2</sup> Während bei Texten der gleichen Autorin / des gleichen Autors trotz aller Veränderungen im Laufe der Schaffensperiode und trotz einer möglichen Intention auf wechselnde Stilformen doch von einem mit sich selbst identischen Subjekt der Text-Origo ausgegangen werden kann, kann man Texte einer Gattungen nur schwerlich einer gemeinsamen Origo-Instanz zuordnen. Zudem handelt es sich bei der Zugehörigkeit eines Textes zu einer Gattung nicht um ein objektives Faktum, sondern um ein Attributionsphänomen – Literaturwissenschaftler\*innen schreiben Texten die Eigenschaft zu, einer Gattung anzugehören. Solche Zuweisungen sind nicht immer eindeutig möglich, da nicht wenige Texte im Spannungsfeld zwischen mindestens zwei Gattungen verortet werden können;<sup>3</sup> so zeichnet sich etwa der *Willehalm* von Wolfram von Eschenbach durch Gattungsinterferenzen aus, in dem neben dem Überlebenskampfmotiv (ein Kennzeichen der *Heldenepik*) auch zahlreiche höfische Passagen (Kennzeichen *höfischer Epik*) vorkommen. Mitunter folgen Gattungszuweisungen auch pragmatischen Kriterien, wenn etwa der *Willehalm* für die Zwecke der Prüfungen im bayerischen Staatsexamen rubriziert werden muss und in diesem Rahmen das vereindeutigende Etikett »Heldenepik« erhält. Dazu kommt, dass Gattungen nicht historisch konstant und gleichförmig bleiben – Gabriel Viehhauser zeigt in seiner Studie zum mittelalterlichen Minnesang, wie sich Gattungswandel auch digital abbilden lässt.<sup>4</sup>

In Studien zur Autorschaftsattributions wurden für schwierige Fälle unklarer Autorschaft (etwa bei sehr kurzen und nicht-normalisierten mittelhochdeutschen Texten) Optimierungsverfahren entwickelt – etwa das *Gute-Wörter-Verfahren*, bei dem nicht alle *Most-Frequent-Words* (MFW), sondern signifikante Wortformen verwendet werden.<sup>5</sup> In der vorliegenden Studie soll geprüft werden, ob sich eine Verbesserung der Erkennungsleistung mit Hilfe des Guten-Wörter-Verfahrens auch bei Gattungsfragen einstellt. Untersucht werden 100 deutsche Texte, die aus dem 19. Jahrhundert stammen oder die kurz davor bzw. danach verfasst wurden. Damit das *Gattungssignal* möglichst zielgerichtet und unbeeinflusst durch *Autorsignale* oder *Übersetzersignale* untersucht werden kann, wird vermieden, mehrere Texte der gleichen Autor\*innen und ins Deutsche übersetzte Texte zu berücksichtigen.

---

<sup>1</sup> Vgl. etwa Büttner et al. 2017.

<sup>2</sup> Vgl. Hettinger et al. 2016a, S. 158. Hettinger et al. 2015 berichten über eine verbesserte Erkennungsleistung mit Hilfe von SVMs, die auf der Basis von LDA-Topics erzielt wurde; vgl. ähnlich Hettinger et al. 2016b. Kim et al. 2017 vergleichen englische Abenteuerromane, humoristische Romane, Science Fiction, Liebesgeschichten und Detektivromane auf der Grundlage von MFW-Bag-of-Words (Baseline), »emotional arcs« und einem lexikalischen Emotionsmodell mit Hilfe von Maschinellem Lernen (RF und MLP). Vgl. weiterhin Schöch 2020; Calvo Tello 2019; Underwood 2016; Ardanuy / Sporleder 2014; Underwood et al. 2013; Eder / Rybicki 2011; Sharoff et al. 2010; Stamatatos et al. 2000; Kessler et al. 1997.

<sup>3</sup> Vgl. zu Gattungshybriden etwa Fuchs 1997; Schulz 2000.

<sup>4</sup> Vgl. Viehhauser 2017.

<sup>5</sup> Zur Verbesserung durch »Gute Wörter« bei Autorschaftsfragen vgl. Dimpel / Proisl 2019.

Diese Studie ist also auf eine technische Fragestellung begrenzt – auf den Beitrag des Gute-Wörter-Verfahrens zu einer verbesserten automatischen Textsortenerkennung. Angestrebt wird nicht, traditionelle literaturwissenschaftliche Genre-Bestimmungen zu kritisieren oder zu präzisieren.<sup>6</sup> Da versucht wird, eine Textsortenerkennung auf lexikalischer Basis vorzunehmen, könnte dieser Versuch als Modellierung<sup>7</sup> einer Unterscheidbarkeit von Textsorten beschrieben werden. Für Computermodelle ist wie auch bei herkömmlichen Modellen das Merkmal der verkürzten Repräsentation wesentlich – das modellierte Objekt wird nicht vollständig durch das Modell abgebildet.<sup>8</sup> Aufgrund dieser Verkürzung ist es in der Regel problematisch, davon zu sprechen, dass sich Ergebnisdaten, die auf der Basis von digitalen Modellen gewonnen werden, unmittelbar dafür eignen, etwa hermeneutische Thesen zu verifizieren oder zu falsifizieren.<sup>9</sup>

Gegenüber Studien, die auf Black-Box-Verfahren wie Maschinelles Lernen (siehe Anmerkung 2) setzen, haben Burrows' Delta und das Gute-Wörter-Verfahren den Vorteil, dass die Berechnungsgrundlage transparent nachvollzogen werden kann. Zudem lässt sich überprüfen, welche Wortformen zur Textsortenunterscheidung besonders gut beitragen (siehe Anhang, Tabelle 15).

## 2. Korpusgestaltung und Präprocessing

Verwendet wurden Texte, die im Internet frei verfügbar sind. Die meisten Texte entstammen dem Textgrid-Repository. Die Texte wurden dann als Abenteuerroman, Bildungsroman, Gesellschaftsroman, Komödie oder Tragödie eingestuft, wenn in einer literaturgeschichtlichen Darstellung oder in einem Forschungsbeitrag eine einschlägige Bezeichnung vorgefunden wurde.<sup>10</sup>

Während der Einfluss des Übersetzersignals noch weniger gut erforscht ist,<sup>11</sup> kann das Autorsignal als ein starkes stilometrisches Signal gelten.<sup>12</sup> Falls beispielsweise bei der Untersuchung von Romansubgenres zahlreiche Texte von Karl May im Korpus vorhanden sind, ist denkbar, dass die Erkennung des Subgenres Abenteuerroman durch das Autorsignal von Karl May positiv beeinflusst wird – bei nicht wenigen Autor\*innen ist eine gewisse Präferenz für eine oder wenige Gattungen erkennbar. Umgekehrt sind auch Fälle denkbar, in denen die gleichen Autor\*innen in verschiedenen Gattungen wirken, so dass ihre Texte aufgrund des Autorsignals zusammenclustern, obwohl sie verschiedenen Gattungen zugeschrieben werden. Hettinger et al. berichten von einem Sinken der Erkennungsleistung, wenn man Autorduplikate aus dem zuvor examinierten Korpus herausnimmt.<sup>13</sup>

Ein Problem bei der Korpus-Zusammenstellung sind Gattungsinterferenzen und mehrfache Labels: So ist Wilhelm Raabes *Abu Telfan oder Die Heimkehr vom Mondgebirge* sowohl als Entwicklungsroman, Gesellschaftsroman, Abenteuerroman, Bildungsroman als auch Zeitroman eingestuft worden. Rolf Selbmann etwa bespricht einige offenbar als prototypisch verstandene Bildungsromane unter der Überschrift »Zwischen Individualroman und Gesellschaftsroman«<sup>14</sup>; andernorts beruft sich Selbmann auf Benno von Wiese, der »die *Epigonen* [Immermann] zugleich als »Entwicklungsroman«, als »Abenteuerroman«, als »modernen Roman«, wie auch als »gesellschaftlichen Zeitroman« versteht.<sup>15</sup>

Bei der Korpus-Erstellung wurden Texte vermieden, die beispielsweise sowohl als Gesellschaftsroman als auch als Bildungsroman bezeichnet wurden, wodurch sich die Zahl der verfügbaren Texte deutlich reduziert hat. Weitere Einschränkungen ergaben sich durch das Vermeiden von Übersetzungen und Autorenduplikaten. Nur mit einiger Mühe war es möglich, je Textsorte 20 digitale Texte zu finden, die diese drei Kriterien erfüllt haben. Weiterhin wurden stark dialektal geprägte Texte wie »Sozialaristokraten« von Arno Holz nicht ins Korpus genommen. Selbstverständlich wäre es wünschenswert, diese Tests auf einer breiteren Textgrundlage wiederholen zu können.

<sup>6</sup> Gittel / Köppe 2022, S. 13–22, kritisieren die Studie von Underwood 2016 für ihre Thesenbildung zu Genre-Grenzen, dem Generationen-Bezug von Genres und der Kohärenz von Genres, die auf der Grundlage von linguistischen Parametern mit Hilfe von maschinellem Lernen erfolgt. U.a. anhand der Textsorten Pastiche und Parodie stellen Gittel / Köppe in Frage, inwieweit linguistische Textmerkmale hinreichend für eine Genre-Bestimmung sein können.

<sup>7</sup> Zum Modellieren als zentrale Tätigkeit im DH-Bereich vgl. McCarty 2005.

<sup>8</sup> Stachowiak 1973, S. 132. Vgl. zur Approximation bei der Modellbildung Saam / Gautschi 2015, S. 26–38. McCarty 2005, S. 24, weist darauf hin, dass auch in der traditionellen Literaturwissenschaft Modelle omnipräsent sind – etwa bei der Beschreibung von Epochen. Gittel / Köppe 2022, S. 20, kritisieren insbesondere, dass die digitale Modellierung von Underwood anders als literaturwissenschaftliche Beschreibungen Kontexte und paratextuelle Informationen nicht einbeziehen, diese können jedoch für die Erkennung der Gattung eines Textes entscheidend sein.

<sup>9</sup> Zur Trennung von Ergebnisdaten und Interpretation vgl. auch Dimpel 2015.

<sup>10</sup> Dieses Verfahren lässt sich durchaus kritisieren: Bedacht wird dabei nicht, wie oft einem Text die Eigenschaft zugesprochen wird, zu einer Textsorte zu gehören. Unberücksichtigt bleibt auch, ob die Zuordnungen auf einheitlichen Genre-Definitionen basieren. Um eigene Textsortenmodelle zu entwickeln und die Zuordnungen auf dieser Basis zu überprüfen, standen für die Studie nicht die nötigen Ressourcen zur Verfügung. Insoweit deviante Epochenbegriffe eingehen sollten, würden damit allerdings gelebte Praktiken im Fach berücksichtigt.

<sup>11</sup> Vgl. Büttner / Proisl 2016.

<sup>12</sup> Vgl. Schöch 2014.

<sup>13</sup> Hettinger et al. 2016a, S. 161.

<sup>14</sup> Vgl. Selbmann 1994, S. 96–120.

<sup>15</sup> Selbmann 1994, S. 111.

Für die Evaluierung des Gute-Wörter-Verfahrens wurden zwei überschneidungsfreie Teilkorpora verwendet: Die 50 Texte der Ermittlungsgruppe, auf deren Grundlage die Gute-Wörter-Listen berechnet werden, sind nicht enthalten in der Kontrollgruppe (ebenfalls 50 Texte), die die Qualität der Gattungserkennung erfasst.

Autorduplikate im Korpus haben sich zwar nicht ganz vermeiden lassen, aber es konnten doch Vorkehrungen getroffen werden, dass Autorduplikate weder bei der Berechnung der Guten Wörter noch bei der Evaluation im jeweiligen Test berücksichtigt wurden. Doppelte Autor\*innen, die jeweils einmal in der Kontrollgruppe und einmal in der Ermittlungsgruppe vorhanden sind, sind unproblematisch. Sichergestellt ist zudem, dass innerhalb einer Textsorte in den jeweils zehn Texten der Ermittlungs- und Kontrollgruppe kein Autorenduplikat vorkommt. Zudem wurden in den Fällen, in denen sich doppelte Autor\*innen innerhalb der Kontroll- bzw. Ermittlungsgruppe nicht ganz vermeiden lassen, Texte der Duplikat-Autor\*innen nur als *Ratetext* und nie als Vergleichstext im Vergleichskorpus (dazu mehr im [folgenden Abschnitt](#)) verwendet, so dass in jedem einzelnen Testlauf ausschließlich Texte verschiedener Autor\*innen verwendet wurden.

Im Vorfeld der Tests wurden einige Präprocessing-Schritte unternommen. Bei den Dramen habe ich die Regieanweisungen und die Sprecher\*innenangaben entfernt. Die Zeichensätze wurden nach *ANSI* vereinheitlicht, Sonderzeichen mit Ausnahme der deutschen Umlaute wurden vereinheitlicht, Groß- in Kleinbuchstaben konvertiert, Zahlen eliminiert. Weiterhin wurden die ersten 10 % der *Token* entfernt – mit diesem verbreiteten Verfahren werden paratextuelle Informationen und Besonderheiten am Textanfang beseitigt.

### 3. Gute Wörter berechnen – Ermittlungsgruppe

Das Verfahren zur Ermittlung der Guten Wörter ist ausführlich dokumentiert.<sup>16</sup> Für das Setting ist elementar, dass ein Text als *Ratetext* verwendet wird und gegen ein Vergleichskorpus mit meist 15 bis 30 *Distraktortexten* getestet wird. Das Vergleichskorpus enthält jedoch auch einen Vergleichstext der Zielklasse – bei Autorschaftsfragen ist also ein Text von der Autorin / dem Autor im Vergleichskorpus, von der / dem auch der *Ratetext* stammt; bei Gattungsfragen ein Vergleichstext der gleichen Gattung.

Wie bei Burrows' Delta üblich, wird für jedes Wort der *Most-Frequent-Words* (MFWs) die relative Häufigkeit gezählt, Standardabweichung und *Z-Werte* berechnet und sodann die *Z-Wert-Differenz* zwischen dem *Ratetext* und jedem Vergleichstext. Zentral für die Ermittlung der Guten Wörter sind die *Level-2-Differenzen*, die man berechnet als Differenz aus der *Z-Wert-Differenz* zwischen *Ratetext* und *Distraktortext* einerseits und der *Z-Wert-Differenz* zwischen *Ratetext* und dem Vergleichstext der Zielklasse andererseits. Auf positiven *Level-2-Differenzen* beruht eine funktionierende Erkennung der Zielklasse. Negative *Level-2-Differenzen* sind ein Störfaktor für die Erkennung der Zielklasse.

In einem Setting mit nur einem *Distraktortext* und zwei Texten der gleichen Klasse ist mathematisch unmittelbar evident, dass Wörter mit positiver *Level-2-Differenz* zu einem niedrigen *Delta-Wert* beitragen. In einem größeren Setting mit mehreren *Distraktortexten* sind verschiedene Parameter denkbar, mit deren Hilfe die Liste der Guten Wörter erstellt werden kann. Dimpel / Proisl haben gezeigt, dass *Parametersets* mit einem *Spitzenwertkriterium* zwar eine besonders gute Leistung bei Autorschaftserkennung erbringen, jedoch auch so viele *False-Positives* produzieren, dass dieses Parameterset problematisch ist.<sup>17</sup>

Verwendet wird für jede Textsorte nun eine Liste mit den Wortformen der durchschnittlich höchsten *Level-2-Differenzen* von allen *Ratetexten* zu allen *Distraktortexten*. Um diese Liste der Mittelwerte an hohen *Level-2-Differenzen* zu erstellen, wird jeweils einer von zehn Texten der Zielgattung ins *Distraktorkorpus* als Gattungsvergleichstext gegeben. Die neun anderen Texte der Ermittlungsgruppe der jeweiligen Gattung werden reihum als *Ratetext* verwendet. Zu dem *Ratetext*, dem Gattungsvergleichstext und je einem der *Distraktortexte* wird die *Level-2-Differenz* berechnet. Aus diesen *Level-2-Differenzen* wird der Mittelwert der *Level-2-Differenzen* für diesen *Ratetext* und diesen Gattungsvergleichstext zu allen 20 *Distraktortexten* gebildet. Bei einem Gattungsvergleichstext und neun *Ratetexten* fallen für jede Wortform neun durchschnittliche *Level-2-Differenzen* an. Dieses Verfahren wird zehnmal wiederholt, so dass reihum jeder Text der Ermittlungsgruppe als Gattungsvergleichstext ins *Distraktorkorpus* gegeben wird und die anderen neun Texte als *Ratetexte* »gegen« diesen getestet werden. Es fallen also insgesamt pro Wortform  $20 \times 9 \times 10$  *Level-2-Differenzen* an, aus denen schließlich ein weiterer Mittelwert gebildet wird. Dieses Verfahren wird für jede Textsorte durchgeführt, es fallen also fünf textsortenspezifische Listen mit Guten Wörtern an.

---

<sup>16</sup> Dimpel 2018a; Dimpel et al. 2019; vgl. weiterhin Dimpel 2018b. Ein didaktisch aufbereiteter Foliensatz steht [hier](#).

<sup>17</sup> In Dimpel / Proisl 2019.

Im Distraktorkorpus befinden sich für jede der vier Textsorten der Nicht-Zielklasse jeweils die Ermittlungsgruppentexte mit Nummern 01–05. Da für die wenigen Autorduplikate im Ermittlungsgruppenkorpus hohe Nummern (08, 09) vergeben wurden, ist bei Bildung der Gute-Wörter-Listen kein Autoduplikat im Spiel.<sup>18</sup>

Zudem soll vermieden werden, dass Wortformen, die in den Ratetexten – also innerhalb der Zielgattung – recht selten vorkommen, berücksichtigt werden. Damit eine Wortform bei der Bildung der Liste der Guten Wörter berücksichtigt wird, muss sie in mindestens vier von neun Ratetexten vorkommen. Damit sollen Eigenheiten von Einzeltexten, die mutmaßlich weniger relevant für die Gattung sind, unberücksichtigt bleiben. Dass es sich bei dem Parameter >4 von 9< um einen geeigneten Parameter handelt, wurde in Prätests mit kleinem Korpus und niedriger Iterationszahl ermittelt.<sup>19</sup>

## 4. Evaluierung – Kontrollgruppe

Die fünf Listen der Guten Wörter werden in vier textsortenbezogenen Kombinationen mit Texten der Kontrollgruppe evaluiert:

- Test A) Fünf Textsorten: ABE, BIL, GES, KOM, TRA<sup>20</sup>
- Test B) Drei Textsorten: ABE, BIL, KOM
- Test C) Drei Textsorten: ABE, KOM, TRA (ohne verschiedene Roman-Subgenres)
- Test D) Drei Textsorten: ABE, BIL, GES (ausschließlich Roman-Subgenres)

Da die Unterscheidung von Bildungs- und Gesellschaftsroman aufgrund der thematischen Nähe beider Subgenres eine besondere Herausforderung darstellt, ist für die Testreihen B und C die beste Unterscheidungsleistung zu erwarten.

Für die Testreihen A und D wird angelehnt an Studien zu mittelhochdeutschen Texten<sup>21</sup> zunächst ein reiner *Erkennungsquotentest* mit fünf Vergleichstexten der Zielklasse durchgeführt; für alle vier Testreihen wird ein ARI-Test (*Adjusted Rand Index*) durchgeführt, bei dem zusätzlich auch die Erkennungsquoten (*Recall*), False-Positives und *F1-Werte* ausgegeben werden – zum Setting siehe unten.

Die kürzeste Liste der Guten Wörter, die alle Wortformen mit einer Level-2-Differenz von >0,2 enthält, umfasst bei den Komödien 495 Wortformen, die längste Liste bei den Abenteuerromanen 637 Einträge. Eine Level-2-Differenz von >0,4 ist bei den Komödien bei den Wortformen mit den Nummern 1–254 vorhanden, bei den Abenteuerromanen bei den Wortformen 1–189. Auf einen Test, der exakt die in Dimpel / Proisl 2019 geprüften Schwellenwerte ermittelt, wird verzichtet; getestet wird vielmehr mit 200, 300 und 400 MFWs. Wenn die Guten Wörter nicht in ausreichend vielen Texten im aktuellen Test vorhanden sind,<sup>22</sup> wird das Gute Wort nicht verwendet. Wenn dadurch nicht mehr ausreichend viele Gute Wörter vorhanden sind, wird die MFW-Liste im jeweiligen Test mit herkömmlichen MFWs ergänzt. Es werden also nicht unbedingt ausschließlich Gute Wörter berücksichtigt; insofern ist im Folgenden auch von einer *bevorzugten Verwendung der Guten Wörter* die Rede.

Näherungsweise bildet ein Test mit 200 MFWs einen Level-2-Differenzen-Mittelwert >0,4 und ein Test mit 300 MFWs einen Level-2-Differenzen-Mittelwert >0,2 ab. In der Liste für die Gesellschaftsromane – sie liegt hinsichtlich ihrer Länge im Mittelfeld – ist bei Wortform Nr. 300 eine Level-2-Differenz von 0,32 vorhanden.

<sup>18</sup> Weitere Parameter für die Ermittlung der Guten Wörter: Verwendet wurden volle Texte nach Entfernung der ersten 10 % der Wortformen. Die häufigsten 1.200 MFWs wurden verwendet. Experimente mit 1.500 MFWs haben schlechtere Ergebnisse hervorgebracht. Dies hängt vermutlich damit zusammen, dass die Komödien und Tragödien teils recht kurz sind. Der kürzeste Text kommt nach dem Entfernen der ersten 10 % auf 5.473 Wortformen. Aus Rechenzeitgründen wurden Wortformen nach 75.000 Wortformen nicht mehr berücksichtigt (Cutoff) – über die Hälfte der Texte ist ohnehin nicht länger als 50.000 Wortformen.

<sup>19</sup> In einem weiteren Prätest wurde zunächst versucht, jeweils fünf Texte der Ermittlungsgruppe in einen Pseudo-Gattungstext zusammen zu kopieren (mit Cutoff bei 75.000 Wortformen) und diese Datei als Vergleichstext der Zielklasse im Vergleichskorpus zu verwenden. >Gegen< dieses Vergleichskorpus wurden einzeln die übrigen fünf Texte der Ermittlungsgruppe als Ratetexte getestet. Die Gute-Wörter-Listen, die in diesem Verfahren erzeugt wurden, haben ebenfalls schlechtere Ergebnisse hervorgebracht als die Listen, die im oben beschriebenen >Reihum<-Verfahren generiert wurden.

<sup>20</sup> Abkürzungen: ABE: Abenteuerroman, BIL: Bildungsroman, GES: Gesellschaftsroman, KOM: Komödie, TRA: Tragödie.

<sup>21</sup> Vgl. etwa Büttner et al. 2017.

<sup>22</sup> Weiterhin werden von der MFW-Liste nur Wortformen verwendet, die in mindestens zwei verschiedenen Texten des Korpus vorkommen. Theoretisch denkbar ist, dass in einem Text beispielsweise ein Figurenname derart hochfrequent vorkommt, dass er in die Liste der 400 häufigsten Wörter gelangt.

## 4.1 Setting: Erkennungsquotentest mit fünf Vergleichstexten der Zielklasse

Bei dieser Evaluierung kommen fast alle 50 Texte der Kontrollgruppe zum Einsatz. In das Vergleichskorpus werden jeweils fünf (von zehn in der Kontrollgruppe vorhandenen) Texte pro Gattung per Losverfahren gegeben, die nicht zu der Zielgattung, für die jeweils auch die Guten Wörter berechnet wurden, gehören. Bei der Testreihe A) werden also für vier Gattungen je fünf Distraktortexte ausgelost, insgesamt damit 20 Distraktortexte.

Wenn ein Text als Distraktortext gelost wird, zu dem ein Autorenduplikat in den Kontrollgruppentexten der Zielgattung vorhanden ist, dann wird dieser Distraktortext zurückgelegt; stattdessen wird ein anderer Distraktortext dieser Gattung verwendet. Es ist also sichergestellt, dass die drei Textpaare, deren Autor\*innen in den Kontrollgruppentexten doppelt vertreten sind, nur als Ratetexte und nicht als Distraktortexte im Vergleichskorpus berücksichtigt werden, so dass auch hier Autorenduplikate das Gattungssignal nicht überlagern können.<sup>23</sup>

Bei der Zielgattung wird zunächst reihum jeweils einer der zehn Kontrollgruppentexten als Vergleichstext ins Vergleichskorpus gegeben. Vier weitere Texte der Zielgattung werden jeweils ebenfalls als Vergleichstexte dem Vergleichskorpus zugelost. Die übrigen fünf Kontrollgruppentexte der Zielgattung werden als Ratetexte verwendet.

Wenn dieser Test mit Volltexten (gekürzt auf 100.000 Wortformen) durchgeführt wird, werden pro Gattung zehn Durchgänge absolviert, um die Zufälligkeiten bei der Auslosung auszugleichen; verwendet werden die Durchschnittswerte aller Durchgänge. Wenn bei den Tests das *Bag-of-Words-Verfahren* zum Einsatz kommt, wird die Textzusammenstellung für jeweils 200 Bag-of-Words pro Gattung neu ausgelost. Als Bag-of-Words-Größe wird 10.000 Wortformen angesetzt. Standardmodus ist ›Ziehen ohne Zurücklegen‹. Wenn ein Text – wie bei den kürzeren Komödien und Tragödien – weniger als 11.000 Wortformen umfasst, gilt für diesen Text der Modus ›Ziehen mit Zurücklegen‹.

Insgesamt befinden sich fünf Vergleichstexte der Zielgattung und 20 Distraktortexte (bei Test A) bzw. zehn Distraktortexte (bei den Tests B–D) im Vergleichskorpus. Die erwartete Erkennungsquote bei einer Zufallsverteilung liegt damit bei 20 % (A) bzw. bei 33 % (B–D).

Da in der vorliegenden Studie überprüft werden soll, ob und inwieweit die bevorzugte Berücksichtigung der Guten Wörter zu einer verbesserten Textsortenerkennung führt, wird als Baseline im jeweiligen Test das gewählte Verfahren ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung betrachtet. Eine allgemeine Baseline kann nicht angegeben werden: Einige der Studien, die in Fußnote 2 genannt sind, kommen zu F1-Werten etwas über 0,8, manche kommen zu etwas höheren, andere auch teils zu deutlich niedrigeren Ergebnissen. Allerdings sind die Studien nicht vergleichbar: Verwendet werden verschiedene Korpora, verschiedene Sprachen, teils übersetzte Texte, verschiedene Genres bzw. Subgenres, teils auch nicht-literarische Texte, verschiedene Analyseverfahren und Auswertungsmethoden. Der Umgang mit Autorduplikaten ist ebenso wenig einheitlich wie der Umgang mit mehrfachen Gattungslabels.<sup>24</sup>

## 4.2 Setting: Tests mit F1-Wert und ARI

Bei diesem Testverfahren gebe ich jeweils alle zehn Kontrollgruppentexte der Nicht-Zielgattungen als Distraktortexte ins Korpus – es sei denn, es befindet sich ein Autorenduplikat zu einem Text der Zielgattung darunter; in diesem Fall wird dieser Distraktortext für den Test zur jeweiligen Zielgattung ersatzlos aus dem Korpus genommen, so dass statt 40 nur 38 oder 39 Distraktortexte (Testreihe A) oder statt 20 nur 18 oder 19 Distraktortexte (Testreihe B–D) verwendet werden. In einem alternativen Versuch (nur Testreihen A und D) werden nur fünf zufällige Distraktortexte je Nicht-Zielgattung (ohne Autorenduplikate zur Zielgattung) verwendet. Weiterhin werden alle zehn Texte der Zielgattung ins Korpus gegeben. Für alle möglichen Paare von jeweils zwei Texten des Korpus werden die Delta-Abstände berechnet. Die ARI-Berechnung ist als *Zweiklassenspiel* implementiert: Unterschieden wird zwischen der Zugehörigkeit zur Zielklasse und zur Nicht-Zielklasse. Über die Klassenzugehörigkeit entscheidet dabei der niedrigste Delta-Abstand.

---

<sup>23</sup> In der Kontrollgruppe sind drei Duplikat-Paare vorhanden: Jean Paul (›bil\_19, Flegeljahre‹, ›ges\_13, Blumen, Frucht und Dornenstücke‹), Heinrich Laube (›ges\_15, Junges Europa‹, ›tra\_12, Monaldeschi‹) und Friedrich Schiller (›abe\_12, Geisterseher‹, ›tra\_16, Wallensteins Lager‹).

<sup>24</sup> Ardanuy / Sporleder 2014, S. 37, akzeptieren etwa eine Klassifizierung bei mehrfachen Labels als korrekt, wenn die erkannte Klasse zumindest zu einem der Label passt, während in der vorliegenden Studie angestrebt wurde, Texte mit mehrfachen Labels zu meiden. Eine Vergleichbarkeit der Studien leidet – wie so oft im Bereich der Digital Humanities – auch darunter, dass viele Publikationsorgane den Maximalumfang der Beiträge auf derart wenige Seiten einschränken, dass eine Dokumentation von Setting, Parametern etc. nicht ausreichend möglich ist. Solche Seiteneinschränkungen muten vor allem dort, wo Online-Publikationsformate gewählt werden, geradezu absurd an.

Durchgeführt werden diese Tests ohne weitere Optimierungsmaßnahmen wie das Eliminieren von Pronomina, jedoch mit Berücksichtigung der jeweiligen Gute-Wörter-Liste und mit Z-Wert-Begrenzung auf 1,64.<sup>25</sup> Die Z-Wert-Begrenzung wird aufgrund der Annahme verwendet, dass textspezifisches Vokabular, das nicht zugleich gattungsspezifisches Vokabular ist, auf diesem Weg mitunter aussortiert werden könnte; zugleich könnten Nullwerte, die auf fehlenden Wörtern im Einzeltext beruhen, weniger stark auf den Delta-Wert durchschlagen.

Bei der Auswertung ist zu bedenken, dass der ARI nicht direkt mit einer herkömmlichen Erkennungsquote zu vergleichen ist. Bei dem oben beschriebenen Setting würde eine Zufallsverteilung nicht eine Erkennungsquote von 0%, sondern von 20 % bzw. 33% ergeben. Eine Zufallsverteilung beim ARI-Wert ergibt den Wert 0; Clusterergebnisse, die schlechter als eine Zufallsverteilung sind, führen zu negativen ARI-Werten. Dass der ARI-Wert in vergleichbaren Konstellationen unter der Erkennungsquote liegt (wenn man den Einfluss der False-Positives unberücksichtigt lässt), ist bereits durch den abweichenden Wert für die Zufallsverteilung bedingt. Dieser Effekt verringert sich, je mehr die Erkennungsquote gegen 100 % und der ARI-Wert gegen 1 tendiert.

Neben dem ARI wird hier auch Erkennungsquote (Recall) und False-Positives-Quote notiert, auf deren Basis die Precision ermittelt und der F1-Score für die Zielgattungstexte ausgegeben wird. Für die Erkennungsquoten werden nur die Delta-Abstände zwischen den Texten der Zielgattung zu allen Texten im Korpus herangezogen; für die Nicht-Zielgattungstexte wird also keine Erkennungsquote ermittelt – deren Clusterverhalten geht ohnehin in den ARI ein. Bei der False-Positives-Quote werden die Nicht-Zielgattungstexte berücksichtigt, die zur Zielgattung den niedrigsten Delta-Abstand aufweisen. Da hier ein Zielklassentext gegen ein Korpus mit 9 Zielklassentexten und 38–40<sup>26</sup> Distraktortexten (A) bzw. 18–20 Distraktortexten (B–D) getestet wird, würde eine Zufallsverteilung bei ca. 18,4 % (A) bzw. 31 % (B–D) liegen. Die F1-Werte liegen durchwegs deutlich über den ARI-Werten; bei letzteren gehen auch Anzahl und Clusteringverhalten der Distraktortexte ein.

## 5. Ergebnisse

### Testreihe A: ABE, BIL, GES, KOM, TRA

#### A1: Erkennungsquotentest mit fünf Vergleichstexten der Zielklasse

	200 <sup>a</sup>	300 <sup>a</sup>	400 <sup>a</sup>	
Gute <sup>c</sup> & ZWB <sup>d</sup>	57,3 <sup>b</sup>	<b>63,4</b>	62,3	Bag-of-Words 10.000
Gute	50,0	53,4	51,9	
ZWB	48,2	45,4	44,0	
Basis <sup>e</sup>	45,4	42,9	44,7	
Gute & ZWB	51,6	57,2	56,8	Volltexte
Gute	61,2	60,8	<b>65,6</b>	
ZWB	48,4	48,0	50,4	
Basis	46,4	44,0	47,6	
Zufallsquote	20,0	20,0	20,0	
<sup>a</sup> Anzahl der MFWs, die verwendet werden <sup>b</sup> Erkennungsquote in % <sup>c</sup> Gute: Mit Gute-Wörter-Liste <sup>d</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>e</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung				

Tab. 1: Test A1, Erkennungsquotentest mit fünf Vergleichstexten der Zielklasse, Test A1, ABE, BIL, GES, KOM, TRA. Beim Bag-of-Words-Test mit 10.000 MFWs werden die Texte 200 verschiedenen Bag-Sets pro Gattung zugelost und Mittelwerte gebildet.

<sup>25</sup> Vgl. zur Z-Wert-Begrenzung Evert et al. 2016; Dimpel 2018b.

<sup>26</sup> Es sind 40 Distraktortexte, wenn kein Autorduplikat in der Zielklasse vorliegt; ansonsten je nach Szenario ein oder zwei Distraktortexte weniger.

Die Guten Wörter führen zu einer deutlichen Verbesserung der Erkennungsquote. Der Verbesserungseffekt ist am stärksten ausgeprägt beim Bag-of-Words-Verfahren mit Z-Wert-Begrenzung; der höchste Wert insgesamt wird bei Volltexten und ohne Z-Wert-Begrenzung erreicht. Im Vergleich zu Autorschaftsstudien liegen die Quoten deutlich niedriger – dort werden Werte >90 % erreicht, selbst wenn sich nur ein Text der Zielfautorin / des Zielfautors im Vergleichskorpus befindet.<sup>27</sup>

## A2: ARI-Test mit 4 × 10 Distraktortexten

Volltexte	200 <sup>a</sup>	300 <sup>a</sup>	400 <sup>a</sup>
ARI <sup>c</sup> Gute <sup>d</sup> & ZWB <sup>e</sup>	<b>0,34<sup>b</sup></b>	0,28	0,26
ARI Gute	0,28	0,26	0,25
ARI ZWB	0,33	0,31	<b>0,34</b>
ARI Basis <sup>f</sup>	0,25	0,2	0,25
F1 <sup>g</sup> Gute & ZWB	<b>0,68</b>	0,64	0,65
F1 Gute	0,66	0,65	0,66
F1 ZWB	0,66	0,63	0,66
F1 Basis	0,60	0,56	0,60
EQ <sup>h</sup> Gute & ZWB	<b>58</b>	54	46
EQ Gute	56	56	<b>58</b>
EQ ZWB	54	52	43
EQ Basis	48	44	48
FP <sup>i</sup> Gute & ZWB	12,4	13,9	16,0
FP Gute	14,5	15,5	18,1
FP ZWB	10,8	11,9	<b>9,8</b>
FP Basis	12,9	13,0	12,5
Diff <sup>j</sup> ARI	0,09	0,08	0,01
Diff EQ	10,0	10,0	8,0
Diff FP	0,5	-1,0	-3,5
<sup>a</sup> Anzahl der MFWs, die verwendet werden <sup>b</sup> Erkennungsquote in % <sup>c</sup> ARI: Adjusted Rand Index <sup>d</sup> Gute: Mit Gute-Wörter-Liste <sup>e</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>f</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung <sup>g</sup> F1: Precision und Recall kombiniert <sup>h</sup> EQ: Erkennungsquote in % <sup>i</sup> FP: False-Positives-Quote in % (niedriger ist besser) <sup>j</sup> Diff: Differenzen zwischen Werten mit Gute-Wörter-Liste und mit Z-Wert-Begrenzung zum Basiswert			

Tab. 2: Test A2, ARI-Test mit 4 × 10 Distraktortexten, ABE, BIL, GES, KOM, TRA.

Die besten ARI-Werte finden sich bei der Kombination der bevorzugten Verwendung von Guten Wörtern mit der Z-Wert-Begrenzung bei 200 MFWs sowie bei den Werten mit Z-Wert-Begrenzung. Die Guten Wörter begünstigen eine Verbesserung der Erkennungsquote, die Z-Wert-Begrenzung führt zu einer besseren False-Positives-Quote bei einer etwas niedrigeren Erkennungsquote. Bei der Kombination beider Techniken verbessert sich bei 200 MFWs die False-Positives-Quote gegenüber

<sup>27</sup> Vgl. etwa Büttner et al. 2017.



dem Basiswert leicht; die Erkennungsquote bleibt zugleich deutlich besser. Bei 200 MFWs verbessert sich der ARI-Wert um 0,09 deutlich, jedoch insgesamt auf mäßigem Niveau. Bei 300 und 400 MFWs gehen Gute Wörter mit schlechterer Level-2-Differenz ein; zugleich begünstigt ein größerer Vektor eine bessere Erkennung.

Hier ein Blick in die Einzelwerte für die Gattungen bei 200 MFWs mit Guten Wörtern und Z-Wert-Begrenzung:

200 MFWs	ARI <sup>a</sup>	F1 <sup>b</sup>	EQ <sup>c</sup>	FP <sup>d</sup>
ABE	0,64	0,81	70	2,6
BIL	0,28	0,68	60	15,4
GES	0,11	0,50	40	18,4
KOM	0,38	0,71	60	10,0
TRA	0,27	0,68	60	15,8
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> F1: Precision und Recall kombiniert <sup>c</sup> EQ: Erkennungsquote in % <sup>d</sup> FP: False-Positives-Quote in % (niedriger ist besser)				

Tab. 3: Test A2, ARI-Test mit 4 × 10 Distraktortexten, Einzelwerte, ABE, BIL, GES, KOM, TRA.

Der Gesellschaftsroman erweist sich als problematisch – mit niedriger Erkennungsquote und hoher False-Positives-Rate. Auch Bildungsromane und Tragödien zeigen eine hohe False-Positives-Rate. Überraschend niedrig ist die False-Positives-Rate beim Abenteuerroman, der insgesamt recht gut erkannt werden kann.<sup>28</sup>

Die Bag-of-Words-Technik (hier mit 10.000 Wortformen) führt zu einer Verbesserung der Erkennungsquote bei 300 und 400 MFWs, jedoch auch zu mehr False-Positives, so dass die ARI-Werte etwas schlechter sind. Hier nur die Daten mit Guten Wörtern und Z-Wert-Begrenzung für alle fünf Gattungen:

MFWs	200	300	400
ARI <sup>a</sup>	0,28	0,31	0,3
F1 <sup>b</sup>	0,64	0,69	0,69
EQ <sup>c</sup>	53,3	61,2	61,2
FP <sup>d</sup>	13,9	15,0	15,7
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> F1: Precision und Recall kombiniert <sup>c</sup> EQ: Erkennungsquote in % <sup>d</sup> FP: False-Positives-Quote in % (niedriger ist besser)			

Tab. 4: Test A2, ARI-Test mit 4 × 10 Distraktortexten, Bag-of-Words, ABE, BIL, GES, KOM, TRA.

### A3: ARI-Test mit 4 × 5 Distraktortexten

Volltexte	200	300	400
ARI <sup>a</sup> Gute <sup>b</sup> & ZWB <sup>c</sup>	0,31	<b>0,32</b>	0,30
ARI Gute	0,29	0,29	0,28
ARI ZWB	0,24	0,25	0,3

Tab. 5: Test A3, ARI-Test mit 4 × 5 Distraktortexten, ABE, BIL, GES, KOM, TRA.

<sup>28</sup> Eine ähnliche Tendenz beobachten Hettinger et al. 2016a, S. 160.

ARI Basis <sup>d</sup>	0,19	0,16	0,19
F1 <sup>e</sup> Gute & ZWB	0,76	0,76	<b>0,77</b>
F1 Gute	0,76	0,76	0,76
F1 ZWB	0,70	0,70	0,73
F1 Basis	0,65	0,62	0,65
EQ <sup>f</sup> Gute & ZWB	74,00	73,80	76,00
EQ Gute	74,6	74,6	<b>77</b>
EQ ZWB	63,8	64,6	68
EQ Basis	56,8	54,6	57
FP <sup>g</sup> Gute & ZWB	20,20	19,90	22,10
FP Gute	21,8	22,6	24,8
FP ZWB	19,5	20,3	<b>18</b>
FP Basis	18,9	20,7	19,4
Diff <sup>h</sup> ARI	0,12	0,16	0,11
Diff EQ	17,20	19,20	19,00
Diff FP	-1,30	0,80	-2,70
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> Gute: Mit Gute-Wörter-Liste <sup>c</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>d</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung <sup>e</sup> F1: Precision und Recall kombiniert <sup>f</sup> EQ: Erkennungsquote in % <sup>g</sup> FP: False-Positives-Quote in % (niedriger ist besser) <sup>h</sup> Diff: Differenzen zwischen Werten mit Gute-Wörter-Liste und mit Z-Wert-Begrenzung zum Basiswert			

Tab. 5: Test A3, ARI-Test mit 4 × 5 Distraktortexten, ABE, BIL, GES, KOM, TRA.

Gegenüber dem Test mit 4 × 10 Distraktortexten geht eine Verbesserung der Erkennungsquote mit einer Verschlechterung der False-Positives-Quote einher. Bei 200 MFWs ist die Verschlechterung der False-Positives-Quote nur leicht, bei 400 MFWs deutlich ausgeprägt.

### Test B: ABE, BIL, KOM

Volltexte	200	300	400
ARI <sup>a</sup> Gute <sup>b</sup> & ZWB <sup>c</sup>	0,62	<b>0,68</b>	0,59
ARI Gute	0,51	0,62	0,5
ARI ZWB	0,5	0,46	<b>0,68</b>
ARI Basis <sup>d</sup>	0,5	0,46	0,56
F1 <sup>e</sup> Gute & ZWB	0,88	<b>0,91</b>	0,89
F1 Gute	0,83	0,89	0,86
F1 ZWB	0,81	0,79	0,90
F1 Basis	0,81	0,79	0,84

Tab. 6: Test B, ARI-Test mit 2 × 10 Distraktortexten, ABE, BIL, KOM.

EQ <sup>f</sup> Gute & ZWB	86,7	<b>90,0</b>	<b>90,0</b>
EQ Gute	80,0	<b>90,0</b>	<b>90,0</b>
EQ ZWB	76,7	73,3	86,7
EQ Basis	76,7	73,3	80,0
FP <sup>g</sup> Gute & ZWB	10,0	8,3	13,3
FP Gute	13,3	11,7	20,0
FP ZWB	11,7	13,3	<b>6,7</b>
FP Basis	11,7	13,3	10,0
Diff <sup>h</sup> ARI	0,12	0,22	0,03
Diff EQ	10,0	16,7	10,0
Diff FP	1,7	5,0	-3,3
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> Gute: Mit Gute-Wörter-Liste <sup>c</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>d</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung <sup>e</sup> F1: Precision und Recall kombiniert <sup>f</sup> EQ: Erkennungsquote in % <sup>g</sup> FP: False-Positives-Quote in % (niedriger ist besser) <sup>h</sup> Diff: Differenzen zwischen Werten mit Gute-Wörter-Liste und mit Z-Wert-Begrenzung zum Basiswert			

Tab. 6: Test B, ARI-Test mit 2 × 10 Distraktortexten, ABE, BIL, KOM.

Bei diesem Test bleibt der schwer unterscheidbare Gesellschaftsroman außen vor. Die ARI-Werte verbessern sich in der Zeile »Gute & ZWB« auf ein ordentliches Niveau, der F1-Wert kommt bei 300 MFWs auf ein gutes Niveau. Wiederum führen die Guten Wörter zu besseren Erkennungsquoten und die Z-Wert-Begrenzung zu besseren False-Positives-Quoten. In den Gattungseinzelnwerten (hier nicht abgedruckt) ergibt sich eine optimale Erkennung der Komödie (ARI=1 bei 200–400 MFWs mit Guten Wörtern und Z-Wert-Begrenzung).

### Test C: ABE, KOM, TRA

Hier wird unter den Romansubgenres nur der besser unterscheidbare Abenteuerroman einbezogen. Test C ist der einzige Test in dieser Studie, in der nicht verschiedene Romansubgenres beteiligt sind – hier kann man am ehesten von drei verschiedenen Gattungen sprechen.

Volltexte	200	300	400
ARI <sup>a</sup> Gute <sup>b</sup> & ZWB <sup>c</sup>	0,45	<b>0,53</b>	0,44
ARI Gute	0,47	0,43	0,38
ARI ZWB	0,5	0,45	0,5
ARI Basis <sup>d</sup>	0,45	0,37	0,41
F1 <sup>e</sup> Gute & ZWB	0,78	<b>0,85</b>	0,80
F1 Gute	0,80	0,80	0,75
F1 ZWB	0,81	0,78	0,81
F1 Basis	0,78	0,72	0,75

Tab. 7: Test C, ARI-Test mit 2 × 10 Distraktortexten, ABE, KOM, TRA.

EQ <sup>f</sup> Gute & ZWB	73,33	<b>83,33</b>	80
EQ Gute	76,67	80	73,33
EQ ZWB	76,67	73,33	76,67
EQ Basis	73,33	66,67	70,0
FP <sup>g</sup> Gute & ZWB	13,68	13,68	18,86
FP Gute	15,53	20,7	22,37
FP ZWB	<b>11,93</b>	13,68	12,02
FP Basis	13,68	17,28	15,61
Diff <sup>h</sup> ARI	0,00	0,16	0,03
Diff EQ	0,00	16,66	10,00
Diff FP	0,00	3,60	-3,25
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> Gute: Mit Gute-Wörter-Liste <sup>c</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>d</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung <sup>e</sup> F1: Precision und Recall kombiniert <sup>f</sup> EQ: Erkennungsquote in % <sup>g</sup> FP: False-Positives-Quote in % (niedriger ist besser) <sup>h</sup> Diff: Differenzen zwischen Werten mit Gute-Wörter-Liste und mit Z-Wert-Begrenzung zum Basiswert			

Tab. 7: Test C, ARI-Test mit 2 × 10 Distraktortexten, ABE, KOM, TRA.

Bei den ARI-Werten wird hier ein Niveau erreicht, das zwischen dem ordentlichen Niveau der Testreihe B (ABE, BIL, KOM) und dem mäßigen Niveau der Testreihe A liegt. Wie bislang führt auch hier die Z-Wert-Begrenzung zu einer Verbesserung bei den False-Positives und die Gute-Wörter-Technik zu einer Verbesserung der Erkennungsquote.

Auch hier setzen sich die Durchschnittswerte aus stark schwankenden Einzelwerten zusammen: Während der Abenteuerroman sehr gut clustert, sind die Daten bei den Tragödien ausgesprochen schlecht.

ARI	200	300	400
ABE	1	1	1
KOM	0,26	0,43	0,26
TRA	0,11	0,17	0,06

Tab. 8: Test C, ARI-Test mit 2 × 10 Distraktortexten, Einzelwerte, ABE, KOM, TRA (Volltexte, mit Guten Wörtern und Z-Wert-Begrenzung).

## Test D: ABE, BIL, GES

Anders als in den Testreihen A–C werden hier keine verschiedenen Gattungen, sondern lediglich Romansubgenres untersucht. Dies hat den Vorteil, dass dabei die teils kurzen Komödien und Tragödien gemieden werden können. Das Bag-of-Words-Verfahren kommt hier ohne Zurücklegen aus; ein weiterer Test (D4) mit einem größeren MFW-Bereich wird dadurch möglich.

**D1: Erkennungsquotentest mit fünf Vergleichstexten der Zielklasse**

	200	300	400	
Gute <sup>a</sup> & ZWB <sup>b</sup>	<b>51,2</b>	51,3	<b>61,6</b>	Bag-of-Words 10.000
Gute	50,3	50,9	60,3	
Basis <sup>c</sup>	35,7	44,9	41,2	
Gute & ZWB	45,3	45,3	58,7	Volltexte
Gute	42,7	56,7	49,3	
Basis	34,7	47,3	36,7	
Zufallsquote	33,3			
<sup>a</sup> Gute: Mit Gute-Wörter-Liste <sup>b</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>c</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung				

Tab. 9: Test D1, Erkennungsquotentest mit fünf Vergleichstexten der Zielklasse, ABE, BIL, GES.

Die Erkennungsquote mit den Optimierungstechniken ist beim Bag-of-Words-Verfahren etwas besser als mit Volltexten; bei Test A war jedoch zu beobachten, dass diese Verbesserung mit einer Verschlechterung der False-Positives-Quote einherging. Die Werte sind insgesamt etwas schlechter als die Erkennungsquoten in der folgenden Tabelle beim ARI-Test; die Bag-of-Words-Tests ergeben etwas höhere Werte. Während im ARI-Setting neun Zielklassentexte und 20 Distraktortexte zum Abgleich zur Verfügung stehen, werden hier fünf Zielklassentexte und 10 Distraktortexte verwendet. Die Zufallsquote liegt beim ARI-Setting bei 31%, hier bei 33,3%, also in einer ähnlichen Größenordnung. Als These, die die niedrigeren Werte in diesem Setting erklären könnte, will ich die Überlegung notieren, dass die Gattungserkennung bei einem größeren Korpus besser funktionieren könnte, da hier Einzeltextspezifika weniger Gewicht haben könnten.

**D2: ARI-Test mit 2 × 10 Distraktortexten**

Volltexte	200	300	400
ARI <sup>a</sup> Gute <sup>b</sup> & ZWB <sup>c</sup>	<b>0,22</b>	0,12	0,2
ARI Gute	0,16	0,06	0,12
ARI ZWB	0,06	0,04	0,13
ARI Basis <sup>d</sup>	0,04	0,04	0,07
F1 <sup>e</sup> Gute & ZWB	0,63	0,62	<b>0,67</b>
F1 Gute	0,63	0,54	0,61
F1 ZWB	0,48	0,47	0,57
F1 Basis	0,48	0,47	0,53
EQ <sup>f</sup> Gute & ZWB	55	57,5	<b>65</b>
EQ Gute	56,67	50	60
EQ ZWB	40	40	50
EQ Basis	40	40	46,67
FP <sup>g</sup> Gute & ZWB	<b>18,29</b>	28,55	28,68
FP Gute	24,47	36,58	36,67
FP ZWB	27,98	29,65	26,14
FP Basis	27,98	29,65	27,89

Tab. 10: Test D2, ARI-Test mit 2 × 10 Distraktortexten, ABE, BIL, GES.

Diff <sup>h</sup> ARI	0,18	0,08	0,13
Diff EQ	15,00	17,50	18,33
Diff FP	9,69	1,10	-0,79
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> Gute: Mit Gute-Wörter-Liste <sup>c</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>d</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung <sup>e</sup> F1: Precision und Recall kombiniert <sup>f</sup> EQ: Erkennungsquote in % <sup>g</sup> FP: False-Positives-Quote in % (niedriger ist besser) <sup>h</sup> Diff: Differenzen zwischen Werten mit Gute-Wörter-Liste und mit Z-Wert-Begrenzung zum Basiswert			

Tab. 10: Test D2, ARI-Test mit 2 × 10 Distraktortexten, ABE, BIL, GES.

Wiederum ist, wie ein Blick in die Subgenre-Einzelwerte in der Folgetabelle zeigt, die Erkennung beim Abenteuerroman deutlich besser, das Clustering beim Gesellschaftsroman ist schlechter als eine Zufallsverteilung, es gibt über ein Drittel False-Positives. Damit hängt zusammen, dass das Niveau in der vorausgehenden Tabelle deutlich niedriger ist als bei den Testreihen A und B. Wiederum ist die Z-Wert-Begrenzung für eine Verbesserung bei den False-Positives und die Gute-Wörter-Technik für eine Verbesserung der Erkennungsquote verantwortlich.

200 MFWs	ARI <sup>a</sup>	F1 <sup>b</sup>	EQ <sup>c</sup>	FP <sup>d</sup>
ABE	0,413	0,73	60	5
BIL	0,06	0,57	50	26,3
GES	-0,005	0,54	50	36,8
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> F1: Precision und Recall kombiniert <sup>c</sup> EQ: Erkennungsquote in % <sup>d</sup> FP: False-Positives-Quote in % (niedriger ist besser)				

Tab. 11: Test D2, ARI-Test mit 2 × 10 Distraktortexten, Einzelwerte, ABE, BIL, GES.

### D3: ARI-Test mit 2 × 5 Distraktortexten

Die gleiche Tendenz auf noch schlechterem Niveau zeigt sich bei der Variante mit nur fünf (statt zehn) ausgelosten Distraktortexten je Nicht-Zielklasse:

Volltexte	200	300	400
ARI <sup>a</sup> Gute <sup>b</sup> & ZWB <sup>c</sup>	0,06	0,05	0,05
ARI Gute	0,04	0,05	<b>0,07</b>
ARI ZWB	0,03	0,00	0,04
ARI Basis <sup>d</sup>	-0,01	-0,01	0,02
F1 <sup>e</sup> Gute & ZWB	0,60	0,62	0,64
F1 Gute	0,63	0,61	<b>0,65</b>
F1 ZWB	0,52	0,54	0,59
F1 Basis	0,54	0,55	0,58

Tab. 12: Test D3, ARI-Test mit 2 × 5 Distraktortexten, ABE, BIL, GES.

EQ <sup>f</sup> Gute & ZWB	59,00	65,00	68,30
EQ Gute	67,67	66,00	<b>72,30</b>
EQ ZWB	52,70	55,00	61,30
EQ Basis	55,67	56,33	60,00
FP <sup>g</sup> Gute & ZWB	<b>38,70</b>	46,00	46,00
FP Gute	46,33	50,67	48,70
FP ZWB	49,00	48,00	46,30
FP Basis	49,33	49,33	47,67
Diff <sup>h</sup> ARI	0,07	0,06	0,03
Diff EQ	3,33	8,67	8,30
Diff FP	10,62	3,33	1,67
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> Gute: Mit Gute-Wörter-Liste <sup>c</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>d</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung <sup>e</sup> F1: Precision und Recall kombiniert <sup>f</sup> EQ: Erkennungsquote in % <sup>g</sup> FP: False-Positives-Quote in % (niedriger ist besser) <sup>h</sup> Diff: Differenzen zwischen Werten mit Gute-Wörter-Liste und mit Z-Wert-Begrenzung zum Basiswert			

Tab. 12: Test D3, ARI-Test mit 2 × 5 Distraktortexten, ABE, BIL, GES.

Problematisch an diesem Setting ist die Kombination von zehn Distraktortexten mit neun Zielklassentexten, gegen die im Einzeltest ein Zielklassentext getestet wird. Eine Zufallsverteilung würde eine Erkennungsquote von 31 % ergeben. Die Erkennungsquoten müssten also deutlich höher liegen, um einen guten ARI-Wert zu erzielen. Vor allem ist hier die False-Positives-Rate ausgesprochen schlecht, sie wird auch durch die Z-Wert-Begrenzung nur marginal verbessert. Anders als in den anderen Testreihen sinkt die False-Positives-Rate erst in der Kombination der beiden Optimierungstechniken, allerdings nicht auf ein ordentliches Niveau.

## D4: ARI-Test mit 2 × 10 Distraktortexten und größerem MFW-Bereich

Da die kürzeren Komödien und Tragödien hier unberücksichtigt bleiben, wird es möglich, einen größeren Bereich an MFWs in den Test einzubeziehen. In den übrigen Testreihen werden Listen mit Guten Wörtern verwendet, die mithilfe von 1.200 MFWs ermittelt wurden. Die Anzahl dieser Guten Wörter, deren Level-2-Differenz >0,2 beträgt, liegt dort zwischen 495 und 637 Wortformen. Hier wurden nun die guten Wörter auf der Grundlage von 5.000 MFWs berechnet. Die Anzahl dieser Guten Wörter, deren Level-2-Differenz >0,2 beträgt, liegt hier nun bei 2.572 (ABE), 2.405 (BIL) und 2.530 (GES) Wortformen. Bei der Evaluation werden nun 500–4.000 MFWs verwendet.

Neben den üblichen Tests (in der Folgetabelle von unten nach oben: ›Basis‹: ohne Gute Wörter, ohne Z-Wert-Begrenzung; ›ZWB 1,64‹: nur Z-Wert-Begrenzung, ohne Gute Wörter; ›Gute‹: nur Gute Wörter, ohne Z-Wert-Begrenzung) werden verschiedene Z-Wert-Parameter in Kombination mit den Gute-Wörter-Listen getestet: Bei ›ZWBneg‹ werden positive Z-Werte auf +1,64 und negative Z-Werte auf -0,7 begrenzt, bei ›ZWB 1,0‹, ›ZWB 1,2‹ und ›ZWB 1,64‹ werden wie auch sonst die positiven und die negativen Z-Werte auf den Betrag der angegebenen Werte begrenzt.

Volltexte	500	1000	1500	2000	2500	3000	3500	4000
ARI <sup>a</sup> Gute <sup>b</sup> & ZWB <sup>c</sup> 1,64	0,19	0,19	0,25	0,09	0,09	0,27	0,3	0,22

Tab. 13: Test D4, ARI-Test mit 2 × 10 Distraktortexten und großem MFW-Bereich, Volltexte, ABE, BIL, GES.

ARI Gute & ZWB 1,2	0,17	0,17	0,23	0,17	0,16	<b>0,31</b>	0,28	0,28
ARI Gute & ZWB 1,0	0,14	0,18	0,19	0,21	0,16	0,29	<b>0,33</b>	0,3
ARI Gute & ZWBneg	0,17	0,15	0,22	0,21	0,08	0,3	0,26	0,21
ARI Gute	0,13	0,16	0,22	0,11	0,09	0,23	0,21	0,13
ARI ZWB 1,64	0,13	0,2	0,2	0,18	0,11	0,11	0,1	0,11
ARI Basis <sup>d</sup>	0,08	0,11	0,14	0,14	0,11	0,08	0,09	0,08
F1 <sup>e</sup> Gute & ZWB 1,64	0,68	0,66	0,68	0,65	0,66	0,73	0,72	0,66
F1 Gute & ZWB 1,2	0,67	0,69	0,71	0,68	0,70	0,74	0,73	0,70
F1 Gute & ZWB 1,0	0,66	0,71	0,69	0,70	0,68	0,73	<b>0,76</b>	0,74
F1 Gute & ZWBneg	0,70	0,67	0,69	0,72	0,66	<b>0,75</b>	0,72	0,65
F1 Gute	0,62	0,66	0,66	0,66	0,68	0,71	0,69	0,61
F1 ZWB 1,64	0,57	0,66	0,66	0,66	0,57	0,59	0,56	0,59
F1 Basis	0,51	0,60	0,63	0,63	0,59	0,53	0,56	0,53
EQ <sup>f</sup> Gute & ZWB 1,64	66,67	66,67	66,67	66,67	66,67	70	66,67	60
EQ Gute & ZWB 1,2	63,33	66,67	70	66,67	70	70	70	66,67
EQ Gute & ZWB 1,0	63,33	70	66,67	66,67	66,67	70	<b>73,33</b>	70
EQ Gute & ZWBneg	70	70	66,67	<b>73,33</b>	66,67	<b>73,33</b>	70	60
EQ Gute	60	66,67	66,67	70	73,33	70	66,67	56,67
EQ ZWB 1,64	50	60	60	60	50	53,33	50	53,33
EQ Basis	43,33	53,33	56,67	56,67	53,33	46,67	50	46,67
FP <sup>g</sup> Gute & ZWB 1,64	29,74	35	28,07	38,25	34,74	22,72	19,3	22,81
FP Gute & ZWB 1,2	26,14	27,89	27,98	29,65	31,32	<b>19,21</b>	20,96	22,81
FP Gute & ZWB 1,0	27,81	27,89	27,89	24,39	29,56	20,96	<b>19,21</b>	<b>19,21</b>
FP Gute & ZWBneg	31,32	38,33	26,23	31,49	34,65	22,72	24,47	24,56
FP Gute	33,25	35	35,09	41,84	41,75	26,23	26,23	27,98
FP ZWB 1,64	26,14	20,96	20,96	22,72	24,47	26,23	27,98	27,98
FP Basis	27,89	24,47	22,72	24,47	26,23	29,74	29,74	29,73
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> Gute: Mit Gute-Wörter-Liste <sup>c</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>d</sup> Basis: Ohne Gute-Wörter-Liste und ohne Z-Wert-Begrenzung <sup>e</sup> F1: Precision und Recall kombiniert								

Tab. 13: Test D4, ARI-Test mit 2 × 10 Distraktortexten und großem MFW-Bereich, Volltexte, ABE, BIL, GES.



<sup>f</sup> EQ: Erkennungsquote in %<sup>g</sup> FP: False-Positives-Quote in % (niedriger ist besser)

Tab. 13: Test D4, ARI-Test mit 2 × 10 Distraktortexten und großem MFW-Bereich, Volltexte, ABE, BIL, GES.

Bei den optimalen Parametern (Gute Wörter kombiniert mit Z-Wert-Begrenzung auf 1,0) werden sowohl bei ARI, F1-Score, Erkennungsquote und False-Positives die besten Werte erreicht. Der F1-Score verbessert sich gegenüber dem besten Wert in Testreihe D2 (0,67) nun auf 0,76.

Für die optimalen Werte wurde noch ein Bag-of-Words-Tests durchgeführt mit Bag-of-Words mit je 20.000 Wortformen und 200 Iterationen je Einzelwert (Rechenzeit: gut eine Woche). Die Werte sind hier jedoch wieder schlechter:

BOW 20T	3000	3500
ARI <sup>a</sup> Gute <sup>b</sup> & ZWB <sup>c</sup> 1,0	0,14	0,17
F1 <sup>d</sup> Gute & ZWB 1,0	0,66	0,67
EQ <sup>e</sup> Gute & ZWB 1,0	64,7	<b>64,2</b>
FP <sup>f</sup> Gute & ZWB 1,0	31,01	27,23
<sup>a</sup> ARI: Adjusted Rand Index <sup>b</sup> Gute: Mit Gute-Wörter-Liste <sup>c</sup> ZWB: Mit Z-Wert-Begrenzung auf 1,64 <sup>d</sup> F1: Precision und Recall kombiniert <sup>e</sup> EQ: Erkennungsquote in % <sup>f</sup> FP: False-Positives-Quote in % (niedriger ist besser)		

Tab. 14: Test D5, ARI-Test mit 2 × 10 Distraktortexten und großem MFW-Bereich, Bag-of-Words (20.000 Wortformen), ABE, BIL, GES.

## 6. Fazit

Die Verwendung der Guten Wörter führt zu einer Verbesserung der Erkennungsquoten, die Z-Wert-Begrenzung führt zu einer Verbesserung der False-Positives-Quote. In Kombination führen beide Techniken zu einer Verbesserung der Erkennungsquoten, die nicht auf Kosten einer Verschlechterung der False-Positives-Quote erfolgt – dies ist auch an den verbesserten ARI-Werten ablesbar. Insgesamt bleibt die Gattungserkennung ein schwieriges Geschäft. In Testreihe B wurden bei der Unterscheidung von Abenteuerroman, Bildungsroman und Komödie ordentliche Ergebnisse und zumindest ein guter F1-Wert >0,9 erzielt. Der Test A2 mit allen fünf Textsorten bringt bei 200 MFWs mäßige Erfolge mit F1: 0,68, einer Erkennungsquote von 58 % bei immerhin nur 12,4 % False-Positives hervor (etwas verlagert in Test A3: F1: 0,77, Erkennungsquote: 76%, False-Positives: 22,1%).

In dieser Studie sollte geprüft werden, ob das Gute-Wörter-Verfahren zu einer Verbesserung der Genre-Erkennung beitragen kann. Dazu lässt sich ein positiver Befund festhalten. Wenn man fragt, wie gut die Erkennungsleistung dieser Verfahren bei der Textsortenklassifikation insgesamt ist, ist zu bedenken, dass mit 50 Texten nur ein relativ schmales Korpus evaluiert werden konnte, da Autorduplikate und mehrfache Textsortenlabels vermieden wurden.

Die gewählten Bildungs- und Gesellschaftsromane sowie Tragödien und Komödien auf digitalem Weg zu unterscheiden, bleibt eine anspruchsvolle Herausforderung. Zu überlegen wäre, ob die schlechten Werte beim Gesellschaftsroman damit zusammenhängen könnten, dass gesellschaftliche Zustände auch bei den anderen Textsorten eine wichtige Rolle spielen. Die Unterscheidung des Abenteuerromans von Komödie und Tragödie und die Unterscheidung der Komödie von Abenteuer- und Bildungsroman gelingt in diesem Korpus immerhin fehlerfrei (ARI=1).

## Anhang: Gute-Wörter-Listen

ABE	BIL	GES	KOM	TRA
<ul style="list-style-type: none"> <li>- gang</li> <li>- gilt</li> <li>- herzens</li> <li>- schienen</li> <li>- not</li> <li>- kampf</li> <li>- schlagen</li> <li>- unterbrach</li> <li>- übrigen</li> <li>- schön</li> <li>- o</li> <li>- schwieg</li> <li>- gehn</li> <li>- bisher</li> <li>- ohren</li> <li>- zwischen</li> <li>- geliebten</li> <li>- zukunft</li> <li>- folgte</li> <li>- einsam</li> <li>- geht</li> <li>- name</li> <li>- was</li> <li>- bitte</li> <li>- sehn</li> <li>- ach</li> <li>- ha</li> <li>- ewigen</li> <li>- ward</li> <li>- seltsam</li> <li>- bin</li> <li>- nimmer</li> <li>- dessen</li> <li>- gegen</li> <li>- hinzu</li> <li>- fern</li> <li>- liebe</li> <li>- nun</li> <li>- mein</li> <li>- herz</li> <li>- doch</li> <li>- einen</li> <li>- dank</li> <li>- feind</li> <li>- weh</li> <li>- wars</li> <li>- lust</li> <li>- verzeihen</li> <li>- nimmt</li> <li>- zufall</li> <li>- gehalten</li> <li>- getroffen</li> <li>- tritt</li> <li>- geh</li> <li>- indes</li> <li>- glücklich</li> <li>- eignen</li> <li>- denk</li> <li>- hört</li> <li>- fürstin</li> <li>- macht</li> </ul>	<ul style="list-style-type: none"> <li>- sicherheit</li> <li>- knaben</li> <li>- sorgen</li> <li>- erzählt</li> <li>- knabe</li> <li>- diesmal</li> <li>- erklärte</li> <li>- saßen</li> <li>- doch</li> <li>- seien</li> <li>- weh</li> <li>- soll</li> <li>- war</li> <li>- hier</li> <li>- kennt</li> <li>- nein</li> <li>- ha</li> <li>- in</li> <li>- holen</li> <li>- blieben</li> <li>- ecke</li> <li>- was</li> <li>- halt</li> <li>- stets</li> <li>- legen</li> <li>- sage</li> <li>- wollen</li> <li>- wußte</li> <li>- mirs</li> <li>- ah</li> <li>- geh</li> <li>- gefangen</li> <li>- kommen</li> <li>- geht</li> <li>- wort</li> <li>- bin</li> <li>- frieden</li> <li>- sieh</li> <li>- ja</li> <li>- kenne</li> <li>- will</li> <li>- kommt</li> <li>- gott</li> <li>- fall</li> <li>- konnte</li> <li>- streckte</li> <li>- hast</li> <li>- bitte</li> <li>- oh</li> <li>- ort</li> <li>- müssen</li> <li>- kampf</li> <li>- waffen</li> <li>- rasch</li> <li>- allerdings</li> <li>- laßt</li> <li>- ei</li> <li>- kapitel</li> <li>- waren</li> <li>- unmöglich</li> <li>- sollen</li> </ul>	<ul style="list-style-type: none"> <li>- wißt</li> <li>- zieht</li> <li>- strom</li> <li>- jenem</li> <li>- nase</li> <li>- obgleich</li> <li>- höher</li> <li>- zorn</li> <li>- stimmen</li> <li>- mich</li> <li>- fern</li> <li>- ich</li> <li>- mir</li> <li>- weiber</li> <li>- erklärte</li> <li>- körper</li> <li>- ha</li> <li>- o</li> <li>- hatte</li> <li>- ziel</li> <li>- aufmerksamkeit</li> <li>- schlagen</li> <li>- wars</li> <li>- meinem</li> <li>- setzt</li> <li>- meines</li> <li>- waffen</li> <li>- meiner</li> <li>- meinen</li> <li>- denken</li> <li>- als</li> <li>- offen</li> <li>- hört</li> <li>- hölle</li> <li>- seid</li> <li>- eure</li> <li>- meine</li> <li>- allerlei</li> <li>- machte</li> <li>- ruf</li> <li>- euren</li> <li>- soll</li> <li>- seufzte</li> <li>- eurer</li> <li>- macht</li> <li>- setzte</li> <li>- sich</li> <li>- tische</li> <li>- mein</li> <li>- hielten</li> <li>- gestalten</li> <li>- bin</li> <li>- uns</li> <li>- deine</li> <li>- niemals</li> <li>- hilfe</li> <li>- deinen</li> <li>- sagt</li> <li>- steht</li> <li>- deines</li> <li>- euer</li> </ul>	<ul style="list-style-type: none"> <li>- stieß</li> <li>- tages</li> <li>- schlug</li> <li>- kannte</li> <li>- riß</li> <li>- hing</li> <li>- lag</li> <li>- ergriff</li> <li>- blieben</li> <li>- erschien</li> <li>- flog</li> <li>- standen</li> <li>- hielt</li> <li>- fuhr</li> <li>- stieg</li> <li>- empor</li> <li>- war</li> <li>- und</li> <li>- trat</li> <li>- fiel</li> <li>- weiten</li> <li>- wilden</li> <li>- hatte</li> <li>- wurde</li> <li>- wolken</li> <li>- öffnete</li> <li>- reichte</li> <li>- wenigen</li> <li>- zwischen</li> <li>- dessen</li> <li>- ging</li> <li>- mochte</li> <li>- ist</li> <li>- suchte</li> <li>- lachte</li> <li>- folgte</li> <li>- schien</li> <li>- hob</li> <li>- mannes</li> <li>- stand</li> <li>- ich</li> <li>- wußte</li> <li>- neben</li> <li>- schob</li> <li>- weile</li> <li>- tiefer</li> <li>- stellte</li> <li>- wand</li> <li>- konnte</li> <li>- hörte</li> <li>- blickte</li> <li>- griff</li> <li>- des</li> <li>- sprang</li> <li>- erhob</li> <li>- gespräch</li> <li>- lächeln</li> <li>- mußten</li> <li>- schritte</li> <li>- meer</li> <li>- warf</li> </ul>	<ul style="list-style-type: none"> <li>- mußte</li> <li>- mochte</li> <li>- fuhr</li> <li>- blieben</li> <li>- einigen</li> <li>- hatte</li> <li>- weder</li> <li>- öffnete</li> <li>- hatten</li> <li>- demselben</li> <li>- waren</li> <li>- erzählen</li> <li>- standen</li> <li>- machte</li> <li>- war</li> <li>- wurde</li> <li>- ewig</li> <li>- frieden</li> <li>- unsere</li> <li>- lächelte</li> <li>- sagte</li> <li>- führte</li> <li>- vielmehr</li> <li>- setzte</li> <li>- zeigte</li> <li>- schwere</li> <li>- sieh</li> <li>- wußte</li> <li>- konnte</li> <li>- konnten</li> <li>- schienen</li> <li>- blieb</li> <li>- während</li> <li>- fragte</li> <li>- heraus</li> <li>- schob</li> <li>- gerade</li> <li>- hinzu</li> <li>- kannte</li> <li>- verschwunden</li> <li>- davon</li> <li>- ziemlich</li> <li>- mußten</li> <li>- anderer</li> <li>- erzählte</li> <li>- wandte</li> <li>- mehrere</li> <li>- erkannte</li> <li>- desselben</li> <li>- unterbrach</li> <li>- sprang</li> <li>- begann</li> <li>- ohne</li> <li>- dabei</li> <li>- schüttelte</li> <li>- drückte</li> <li>- erklärte</li> <li>- beiden</li> <li>- endlich</li> <li>- hundert</li> <li>- nachher</li> </ul>

ABE-BIL	51
ABE-GES	43
ABE-KOM	37
ABE-TRA	27
BIL-GES	42
BIL-KOM	32
BIL-TRA	27
GES-KOM	34
GES-TRA	30
KOM-TRA	85

Tab. 16: Duplikate in den Listen der Guten Wörter.

Das Verfahren, dass die Guten Wörter für eine Textsorte in Relation zu den vier anderen Textsorten auf Basis der mehrfach gemittelten Level-2-Differenzen gebildet wurden, bringt es mit sich, dass Wortformen auch dann in eine Gute-Wörter-Liste gelangen können, wenn die Unterscheidungsleistung zu zwei anderen Textsorten nur mäßig, die Unterscheidungsleistung zu zwei nochmals anderen Textsorten jedoch hoch ist. Dadurch ist es möglich, dass einige Wortformen in mehreren gattungsspezifischen Listen auftreten. Die hohe Zahl von 85 Duplikaten bei Komödien und Tragödien ist überraschend; womöglich sind hier viele Wortformen eingegangen, die auf den Unterschieden zwischen Drama und Roman beruhen. Damit korrespondieren könnte auch, dass die Guten Wörter nur mäßig dazu beitragen, die F1-Scores bei der Unterscheidung von Komödie und Tragödie zu verbessern, während die Unterscheidung von Drama und Abenteuerroman fehlerfrei gelingt (vgl. Test C).

Wörter, die man in semantischer Hinsicht vielleicht auch intuitiv mit der Textsorte in Verbindung wollte, sind in den Gute-Wörter-Listen selten – die meisten Wortformen findet man auch sonst in längeren MfW-Listen. Wenn man gezielt sucht, könnten etwa ›fern‹ oder ›Zufall‹ typisch für ein Abenteuer-Sujet sein, ›erklärte‹ für den Bildungsroman (wobei diese Wortform auch bei Gesellschaftsroman und Tragödie vorkommt), ›schwere‹ oder ›verschwunden‹ würden in Tragödien nicht überraschen. Allerdings wäre es keine geringe Herausforderung, Kriterien für eine solche Intuition intersubjektiv nachvollziehbar zu begründen.

Verben stehen recht erwartbar meist in der 3. Person Singular Präteritum, in der Abenteuerroman-Liste sind jedoch relativ viele Verben in der 2. Person Singular Präsens enthalten – womöglich ein Indikator für einen erhöhten Anteil an direkter Figurenrede. Dass ›Berlin‹ in den Listen steht, könnte damit korrespondieren, dass nur zehn Texte je Textsorte für die Berechnung der Listen verwendet wurden; bei einem größeren Korpus würden solche vermutlich textspezifischen Wörter nicht in die Listen eingehen.

## Bibliografische Angaben

- Mariona Coll Ardanuy / Caroline Sporleder: Structure-based Clustering of Novels. In: Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL). Hg. von Association for Computational Linguistics. (EACL 2014, Göteborg, 27.04.2014) Stroudsburg, PA, 2014, S. 31–39. DOI: [10.3115/v1/W14-0905](#)
- Andreas Büttner / Thomas Proisl: Stilometrie interdisziplinär: Merkmalsselektion zur Differenzierung zwischen Übersetzer- und Fachvokabular. In: Modellierung, Vernetzung, Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. DHd 2016. Konferenzabstracts. Hg. von Elisabeth Burr. (DHd: 3, Leipzig, 07.–12.03.2016) Duisburg 2016, S. 70–74. DOI: [10.5281/zenodo.3679331](#) [[Nachweis im GVK](#)]
- Andreas Büttner / Friedrich Michael Dimpel / Stefan Evert / Fotis Jannidis / Steffen Pielström / Thomas Proisl / Isabella Reger / Christof Schöch / Thorsten Vitt: „Delta“ in der stilometrischen Autorschaftsattributions. In: Zeitschrift für digitale Geisteswissenschaften 2 (2017). DOI: [10.17175/2017\\_006](#)
- José Calvo Tello: Gattungserkennung über 500 Jahre. In: DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts. Hg. von Patrick Sahle. (DHd: 6, Frankfurt am Main u. a., 25.–29.03.2019) Frankfurt/Main 2019, S. 292–294. DOI: [10.5281/zenodo.2600812](#)
- Friedrich Michael Dimpel: Der Computerphilologe als Interpret – ein Teilzeit-Empiriker? In: Literatur interpretieren. Interdisziplinäre Beiträge zur Theorie und Praxis. Hg. von Jan Borkowski / Stefan Descher / Felicitas Ferder / Philipp Heine. Münster 2015, S. 339–359. DOI: [10.30965/9783957438973\\_018](#)
- Friedrich Michael Dimpel (2018a): Die guten ins Töpfchen: Zur Anwendbarkeit von Burrows' Delta bei kurzen mittelhochdeutschen Texten nebst eines Attributionstests zu Konrads ‚Halber Birne‘. In: DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts. Hg. von Georg Vogeler. (DHd: 5, Köln, 26.02.–02.03.2018) Köln 2018, S. 168–173. DOI: [10.5281/zenodo.3684897](#)
- Friedrich Michael Dimpel (2018b): Ein Delta-Rätsel: Nicht-normalisierte mittelhochdeutsche Texte, Z-Wert-Begrenzung und ein Normalisierungswörterbuch. Oder: Auf welche Wörter kommt es bei Delta an? Göttingen 2018. (= Dariah-DE Working Papers, 25) URN: [urn:nbn:de:gbv:7-dariah-2017-5-1](#)
- Friedrich Michael Dimpel / Daniel Schlager / Katharina Zeppezauer-Wachauer: Der Streit um die Birne. Autorschafts-Attributionstest mit Burrows' Delta und dessen Optimierung für Kurztexte am Beispiel der ‚Halben Birne‘ des Konrad von Würzburg. In: Digitale Mediävistik. Hg. von Roman Bleier / Franz Fischer / Torsten Hiltmann / Gabriel Viehhauser / Georg Vogeler. Berlin u. a. 2019, S. 71–90. [[Nachweis im GVK](#)]
- Friedrich Michael Dimpel / Thomas Proisl: Gute Wörter für Delta: Verbesserung der Autorschaftsattributions durch autorspezifische distinktive Wörter. In: DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts. Hg. von Patrick Sahle. (DHd: 6, Frankfurt am Main u. a., 25.–29.03.2019) Frankfurt/Main 2019, S. 296–299. DOI: [10.5281/zenodo.2600812](#)
- Maciej Eder / Jan Rybicki: Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? In: Literary and Linguistic Computing 26 (2011), H. 3, S. 315–321. DOI: [10.1093/lc/fqr031](#) [[Nachweis im GVK](#)]
- Stefan Evert / Fotis Jannidis / Friedrich Michael Dimpel / Christof Schöch / Steffen Pielström / Thorsten Vitt / Isabella Reger / Andreas Büttner / Thomas Proisl: Burrows Delta verstehen. In: Modellierung, Vernetzung, Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. DHd 2016. Konferenzabstracts. Hg. von Elisabeth Burr. 2. überarbeitete und erweiterte Ausgabe. (DHd: 3, Leipzig, 07.–12.03.2016) Duisburg 2016, S. 83–86. DOI: [110.5281/zenodo.3679330](#)
- Stephan Fuchs: Hybride Helden: Gwigois und Willehalm. Beiträge zum Heldenbild und zur Poetik des Romans im frühen 13. Jahrhundert. Heidelberg 1997. (= Frankfurter Beiträge zur Germanistik, 31) [[Nachweis im GVK](#)]
- Benjamin Gittel / Tilmann Köppe: On the Distance Between Traditional and DH-Based Genre Theory. In: Digitale Verfahren in der Literaturwissenschaft. Hg. von Jan Horstmann / Frank Fischer. Münster 2022. (= Sonderausgabe Textpraxis. Digitales Journal für Philologie, 6). DOI: [10.17879/64059431694](#)
- Lena Hettinger / Martin Becker / Isabella Reger / Fotis Jannidis / Andreas Hotho: Genre classification on German novels. In: Database and expert systems applications. 26th International Conference. Hg. von Qiming Chen / Abdelkader Hameurlain / Farouk Toumani / Roland Wagner / Hendrik Decker. (DEXA: 26, Valencia, 01.–04.09.2015). Cham u. a. 2015, S. 249–253. DOI: [10.1109/DEXA.2015.62](#) [[Nachweis im GVK](#)]
- Lena Hettinger / Isabella Reger / Fotis Jannidis / Andreas Hotho (2016a): Classification of Literary Subgenres. In: Modellierung, Vernetzung, Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. DHd 2016. Konferenzabstracts. Hg. von Elisabeth Burr. (DHd: 3, Leipzig, 07.–12.03.2016) Duisburg 2016, S. 158–162. DOI: [10.5281/zenodo.3679331](#) [[Nachweis im GVK](#)]
- Lena Hettinger / Fotis Jannidis / Isabella Reger / Andreas Hotho (2016b): Significance Testing for the Classification of Literary Subgenres. In: Digital Humanities 2016. Conference Abstracts. (DH 2016, Krakau, 11.–16.07.2016) Krakau 2016. [[online](#)]
- Brett Kessler / Geoffrey Nunberg / Hinrich Schütze: Automatic Detection of Text Genre. In: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. (ACL: 35 - EACL '97, Madrid, 07.–12.07.1997) Morristown, NJ 1997, S. 32–38. DOI: [10.3115/976909.979622](#) [[Nachweis im GVK](#)]
- Evgeny Kim / Sebastian Padó / Roman Klinger: Investigating the Relationship between Literary Genres and Emotional Plot Development. In: Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Hg. von Beatrice Alex / Stefania Degaetano-Ortlieb / Anna Feldman / Anna Kazantseva / Nils Reiter / Stan Szpakowicz. (SIGHUM: 11, Vancouver, 04.08.2017) Stroudsburg, PA, 2017, S. 17–26. DOI: [10.18653/v1/W17-2203](#)
- Willard McCarty: Humanities Computing. London / New York 2005. [[Nachweis im GVK](#)]
- Nicole J. Saam / Thomas Gautschi: Modellbildung in den Sozialwissenschaften. In: Handbuch Modellbildung und Simulation in den Sozialwissenschaften. Hg. von Norman Braun / Nicole J. Saam. Wiesbaden 2015, S. 15–60. DOI: [10.1007/978-3-658-01164-2](#)
- Christof Schöch: Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In: Literaturwissenschaft im digitalen Medienwandel. Hg. von Christof Schöch / Lars Schneider. Berlin 2014, S. 130–157. (= Philologie im Netz / Beiheft, 7) PDF. [[online](#)]
- Christof Schöch: Computational Genre Analysis. In: Digital Humanities for Literary Studies: Methods, Tools & Practices. Hg. von James O'Sullivan. College Station, TX 2020. Preprint. PDF. [[online](#)]
- Armin Schulz: Poetik des Hybriden. Schema, Variation und intertextuelle Kombinatorik in der Minne- und Aventureepik: ‚Willehalm von Orlens‘ – ‚Partonopier und Meliur‘ – ‚Wilhelm von Österreich‘ – ‚Die schöne Magelone‘. Berlin 2000. (= Philologische Studien und Quellen, 161) [[Nachweis im GVK](#)]
- Rolf Selbmann: Der deutsche Bildungsroman. 2., überarbeitete und erweiterte Auflage. Stuttgart u. a. 1994. (= Sammlung Metzler, 214) [[Nachweis im GVK](#)]
- Serge Sharoff / Zhili Wu / Katja Markert: The Web Library of Babel: evaluating genre collections. In: Proceedings of the 7th International Conference on Language Resources and Evaluation. Hg. von Nicoletta Calzolari / Khalid Choukri / Bente Maegaard / Joseph Mariani / Jan Odijk / Stelios Piperidis / Mike Rosner / Daniel Tapias. (LREC'10: 7, Valetta, 17.–23.05.2010) Paris 2010. PDF. [[online](#)]
- Herbert Stachowiak: Allgemeine Modelltheorie. Wien 1973. [[Nachweis im GVK](#)]
- Efstathios Stamatatos / Nikos Fakotakis / George Kokkinakis: Automatic text categorization in terms of genre and author. In: Computational Linguistics 26 (2000), S. 471–495. DOI: [10.1162/089120100750105920](#)
- Der Streit um die Birne. Autorschafts-Attributionstest mit Burrows' Delta und dessen Optimierung für Kurztexte am Beispiel der ‚Halben Birne‘ des Konrad von Würzburg: Anhang – Dimpel: Gute Wörter und Level-2-Differenzen bei Delta. Hg. von ULB Münster. 2022. [[online](#)]
- Ted Underwood / Michael L. Black / Loretta Auvil / Boris Capitanu: Mapping mutable genres in structurally complex volumes. In: Proceedings of the IEEE International Conference on Big Data. Hg. von Hu Xiaohua. 2 Bde. (Silicon Valley, CA, 06.–09.10.2013) Piscataway, NJ 2013. Bd. 1: S. 95–103. DOI: [10.1109/BigData.2013.6691676](#) [[Nachweis im GVK](#)]
- Ted Underwood: The Life Cycles of Genres. In: Cultural Analytics 2 (2016), H. 2. DOI: [10.22148/16.005](#)

Gabriel Viehhauser: Digitale Gattungsgeschichten. Minnesang zwischen generischer Konstanz und Wende. In: Zeitschrift für digitale Geisteswissenschaften 2 (2017). DOI: [10.17175/2017\\_003](https://doi.org/10.17175/2017_003)

## Tabellenverzeichnis

- Tab. 1: Test A1, Erkennungsquotentest mit fünf Vergleichstexten der Zielklasse, Test A1, ABE, BIL, GES, KOM, TRA. Beim Bag-of-Words-Test mit 10.000 MFWs werden die Texte 200 verschiedenen Bag-Sets pro Gattung zugelost und Mittelwerte gebildet.
- Tab. 2: Test A2, ARI-Test mit  $4 \times 10$  Distraktortexten, ABE, BIL, GES, KOM, TRA.
- Tab. 3: Test A2, ARI-Test mit  $4 \times 10$  Distraktortexten, Einzelwerte, ABE, BIL, GES, KOM, TRA.
- Tab. 4: Test A2, ARI-Test mit  $4 \times 10$  Distraktortexten, Bag-of-Words, ABE, BIL, GES, KOM, TRA.
- Tab. 5: Test A3, ARI-Test mit  $4 \times 5$  Distraktortexten, ABE, BIL, GES, KOM, TRA.
- Tab. 6: Test B, ARI-Test mit  $2 \times 10$  Distraktortexten, ABE, BIL, KOM.
- Tab. 7: Test C, ARI-Test mit  $2 \times 10$  Distraktortexten, ABE, KOM, TRA.
- Tab. 8: Test C, ARI-Test mit  $2 \times 10$  Distraktortexten, Einzelwerte, ABE, KOM, TRA (Volltexte, mit Guten Wörtern und Z-Wert-Begrenzung).
- Tab. 9: Test D1, Erkennungsquotentest mit fünf Vergleichstexten der Zielklasse, ABE, BIL, GES.
- Tab. 10: Test D2, ARI-Test mit  $2 \times 10$  Distraktortexten, ABE, BIL, GES.
- Tab. 11: Test D2, ARI-Test mit  $2 \times 10$  Distraktortexten, Einzelwerte, ABE, BIL, GES.
- Tab. 12: Test D3, ARI-Test mit  $2 \times 5$  Distraktortexten, ABE, BIL, GES.
- Tab. 13: Test D4, ARI-Test mit  $2 \times 10$  Distraktortexten und großem MPW-Bereich, Volltexte, ABE, BIL, GES.
- Tab. 14: Test D5, ARI-Test mit  $2 \times 10$  Distraktortexten und großem MPW-Bereich, Bag-of-Words (20.000 Wortformen), ABE, BIL, GES.
- Tab. 15: Auszug aus den textsortenspezifischen Gute-Wörter-Listen: Jeweils 100 Wortformen mit den höchsten Level-2-Differenzen.
- Tab. 16: Duplikate in den Listen der Guten Wörter.