IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Bigrammhäufigkeiten als Kennzahlen und Evaluationskriterien für die Vermittlung einer verbundenen Handschrift

DEREECHAR-v-[bi|uni]-[KJL|DRC]-2021-10-31-1.0

Technical Report IDS-KL-2021-02

Cyril Belica, Rainer Perkuhn

Leibniz-Institut für Deutsche Sprache, Mannheim

Oktober/November 2021

Inhaltsverzeichnis

1. Einleitung 2
   1.1 Datengrundlage 3
   1.2 Tokenbegriff 4
   1.3 Bigrammzählung 4
   1.4 Unigrammzählung 4
2. Darstellung der Ergebnisse 5
   2.1 Synoptische Darstellung (Unigramme, Top 10 Bigramme je Buchstabe, Top 100 aller Bigramme) 5
   2.2 Visuelle Darstellung der Bigrammverteilung 6
   2.3 Gesamtdarstellung der Bigrammhäufigkeiten 6
3. Vergleichswerte einer breiteren Datengrundlage 7
4. Quellenverzeichnis 8
5. Kontakt 8
6. Lizenzbestimmungen 8

1. Einleitung

In dieser Studie sind korpusbasierte Untersuchungen zu Bigrammhäufigkeiten für eine spezielle Fragestellung dokumentiert.

Die im Rahmen dieser Studie erstellten und über die DEREWO-Webseite zur Verfügung gestellten Materialien sind dafür konzipiert zu helfen, verschiedene Ansätze der Vermittlung einer verbundenen Handschrift zu beurteilen[1]. Vermeintliche Schwierigkeiten bei der Verbindung zweier Buchstaben können eventuell dahingehend relativiert werden, ob diese Buchstabenverbindung quantitativ eine Rolle spielt. Die Ergebnisse dieser Studie sind aber stets vor dem Hintergrund der ausgewählten Datengrundlage und des festgelegten Wortbegriffs einzuordnen.

Bigramme in diesem Sinne sind Zweierfolgen von alphabetischen Zeichen[2]; die Zeichenfolge *Wort* besteht etwa aus den Bigrammen *Wo*, *or* und *rt*. Als Häufigkeit wird für alle Zeichenfolgen berücksichtigt, wie oft diese in einem Korpus belegt sind, d.h. mit wie vielen Token (Instanzen) die Types (Oberflächenformen) vertreten sind. Entsprechend werden die Bigramme kumuliert und mit einer normierten Frequenz *ipm* (Instanzen pro Million) angegeben. Zwecks besserer Vergleichbarkeit mit anderen Angaben oder Angaben anderer Untersuchungen skaliert dieses Maß die Frequenz auf einen relativen Wert, der statt der Summe der beobachteten Ereignisse den fiktiven Wert von 1.000.000 als Summe ansetzt.

Als „Summe der beobachteten Ereignisse“ wird hier die Gesamtanzahl der Buchstaben (Unigramme) der betrachteten Token gesetzt.

Type	**Token-frequenz**
Wort	120
Bart	100
Tor	40

Uni-gramm	**Häufig-keit**
W	120
o	160
r	260
t	220
B	100
a	100
T	40
Summe	1.000

Bi-gramm	**Häufig-keit**	***ipm***
Wo	120	120.000
or	160	160.000
rt	220	220.000
Ba	100	100.000
ar	100	100.000
To	40	40.000

Tab. 1 Grundprinzip der Bigrammzählung und Angabe der normierten Häufigkeit

Die sonst übliche Bezugsgröße eines Korpusumfangs in Anzahl Token erscheint hier nicht angebracht. Die Anzahl der Bigramme steigt zwar mit der Anzahl Token, ist aber auch insbesondere von deren Längen abhängig. Aus einem Token der Länge i werden für diese Untersuchung i-1 Bigramme gebildet. Die durchschnittliche Tokenlänge steigt bei größerem Umfang zwar auf der Vokabularebene deutlich an, auf der Instanzenebene ist der Effekt aber vernachlässigbar. Einen größeren Einfluss hat hier die Zusammensetzung und Aufbereitung (insbesondere natürlich die Tokenisierung) des Korpus: Korpora unterschiedlicher Sprachausschnitte ungefähr gleicher Größe (in „Anzahl Token“) können durchaus stark abweichende durchschnittliche Tokenlängen aufweisen. Da sich dies auch auf die

[1] vgl. Steinig 2017, insbesondere Kap. 6
[2] nicht zu verwechseln mit der sonst auch üblichen Verwendung als Zweiwortfolgen

Bigrammbildungen auswirkt, bleibt als davon unabhängige Größe, den Umfang der betrachteten Daten in Anzahl Zeichen zu messen.

Sei k_i die Anzahl der Token der Länge i, i ≥ 2, da i=1 nicht relevant.

$$\#token = \sum_{i=2}^{max} k_i \qquad \#char = \sum_{i=2}^{max} i * k_i$$

$$\#bigrams = \sum_{i=2}^{max} (i-1) * k_i = \sum_{i=2}^{max} (i * k_i - k_i) = \sum_{i=2}^{max} (i * k_i) - \sum_{i=2}^{max} k_i = \#char - \#token$$

Abb. 1 Zusammenhang zwischen Token-, Bigramm- und Unigrammfrequenzen

Die *ipm*-Werte werden in dieser Studie stets ganzzahlig gerundet angegeben; für die fokussierenden Ergebnispräsentationen (s.u.) wird auf die beiden führenden Ziffern kaufmännisch gerundet.

***ipm* genau**	***ipm* auf Stelligkeit 2 gerundet**
nicht belegt	NA
0-0,49	0
0,5-1,49	1
…	
98,5-99,49	99
99,5-104,99	100
105-114,99	110

***ipm* genau**	***ipm* auf Stelligkeit 2 gerundet**
985-994,99	990
995-1.049,99	1.000
1.050-1.149,99	1.100
…	
9.850-9.949,99	9.900
9.950-10.499,99	10.000
10.500-11.499,99	11.000

Tab. 2 Kaufmännische Rundung der *ipm*-Werte auf Stelligkeit 2

Diese Rundung soll bei der Interpretation davon bewahren, die Genauigkeit der *ipm*-Angabe überzubewerten, da sie zu vielen zufälligen Einflussgrößen unterliegt.

1.1 Datengrundlage

Eine ideale Datengrundlage für diese Untersuchung müsste die produktive Perspektive des Schreibens aus Sicht der Zielgruppe der Handschriftvermittlung wiedergeben. Dazu liegen leider keine aussagekräftigen Daten vor.

Als Annäherung an einen zielgruppen-, insbesondere altersgemäßen Ausschnitt der deutschen Sprache wird das Kinder- und Jugendliteraturkorpus (KJL) des Deutschen Referenzkorpus (DEREKO) des Leibniz-Instituts für Deutsche Sprache (IDS) verwendet, auch wenn dadurch die produktive Perspektive nur indirekt berücksichtigt wird durch das Talent der Autoren und Autorinnen, sich auf altersgemäße Ausdrucksweisen einzustellen.

Das KJL-Korpus in der Fassung der DEREKO Ausgabe 2021-I (Leibniz-Institut für Deutsche Sprache 2021) umfasst 88 Bände mit einem Gesamtumfang von mehr als 4.000.000 Token. Die Bände sind mit Altersempfehlungen ausgezeichnet, z.B. mehr als 30 Bände der Conni-Reihen

des Carlsen-Verlags mit Altersempfehlung ab 8 Jahren, aber auch andere Werke mit höheren Altersangaben wie z.B. Christiane F.s „Kinder vom Bahnhof Zoo“.

Zu diesen Daten wurde eine aktuelle Tokenfrequenzliste mit Werkzeugen der internen KorAP-Infrastruktur[3] erstellt.

1.2 Tokenbegriff

Aus der Menge aller Wortformen (Token) der Frequenzliste wurden nur die berücksichtigt, die entweder mit einem Groß- oder Kleinbuchstaben beginnen und danach nur aus Kleinbuchstaben (jeweils des deutschen Alphabets) bestehen[4]. D.h., dass Wörter mit Bindestrich oder anderen Syngraphemen, mit Binnengroßbuchstaben, aber auch mit Zeichen nicht-deutscher Alphabete (mit Accents, anderen Diakritika oder anderen Zeichen) komplett ignoriert wurden. Bei diesen besteht eher die Gefahr, dass sie die Bigrammverteilungen verzerren, als dass sie neue Information beitragen.

Für die weitere Verarbeitung wurden die Umlaute durch ihr Basiszeichen (Ä → A, ä → a, usw.) ersetzt, da für die handschriftliche Verbindung der Buchstaben lediglich die Basiszeichen relevant sind.

1.3 Bigrammzählung

Auf dieser Datengrundlage und mit diesem Tokenbegriff wurden die Bigrammhäufigkeiten entsprechend ihrer Tokenfrequenz berechnet und für drei Kategorien ausgewiesen: Nachfolgebuchstaben von Großbuchstaben, Nachfolgebuchstaben von Kleinbuchstaben, Vorgängerbuchstaben von Kleinbuchstaben. Die Frequenzen sind in einer Gesamttabelle als ganzzahlig gerundete *ipm* (Instanzen pro Million Wörter), in einer synoptischen Darstellung mit der vorgegebenen Stelligkeit 2 (kaufmännisch gerundet) normiert dargestellt (s.o.).

Besondere Eigenschaften von Bigrammen, etwa Digraph oder Teil eines Trigraphen zu sein, sind bei dieser Vorgehensweise nicht explizit berücksichtigt, da weder Information darüber noch über eine Silbenstruktur oder eine morphematische Segmentierung vorliegen. So können Bigramme gleicher Form an Silben- oder Morphemgrenzen deren Häufigkeiten unterschiedlich stark überlagern[5].

1.4 Unigrammzählung

Für Vergleichszwecke innerhalb dieser Studie werden auch die Häufigkeiten der einzelnen Zeichen (Unigramme) berechnet und veröffentlicht. Zu beachten ist hierbei, dass jedes Zeichen nur einmal berücksichtigt wird, auch wenn es durch die Überlappung im Wortinneren zwei Bigrammen zugeordnet wird: Z.B. fließt das *o* in *Wort* nur einmal mit der Frequenz des Token in die Kumulierung der Unigramme ein, auch wenn es an den beiden Bigrammen *Wo* und *or* beteiligt ist.

Speziell diese Unigrammliste ist aber auch nur vor dem Hintergrund dieser Studie zu interpretieren. Zu allgemeineren Aussagen zu Unigrammen des Deutschen Referenzkorpus sei

---

[3] KorAP (https://korap.ids-mannheim.de/) ist die neue Korpusanalyseplattform der IDS
[4] als regulärer Ausdruck: [A-ZÄÖÜa-zäöüß][a-zäöüß]+
[5] vgl. etwa *ng* in Achtu*ng* vs. a*n*geben

auf eine breiter angelegte Studie verwiesen, deren Ergebnisse ebenfalls auf der DEREWO-Seite publiziert sind (Institut für Deutsche Sprache, 2018).

2. Darstellung der Ergebnisse

Im KJL-Korpus wurden 924 verschiedene Bigramm-Typen mit weit über 15.000.000 Instanzen ermittelt bei einer durchschnittlichen Wortformlänge von 4,99 Zeichen/Wort (genaugenommen: Unigrammtoken pro Wortformtoken, vgl. Tab. 3).

	Anzahl Token	**Anzahl Types**	**Anzahl ignorierte Token (Types)**
Wortformen	4.006.579	105.055	938.378 (5.979)
Unigramme	19.870.142	53	1.040.534
Bigramme	15.863.563	924	

Tab. 3 Zusammenfassung der ausgewerteten bzw. ignorierten Größen im KJL-Korpus

Die Ergebnisse der Auswertung werden auf drei verschiedene Arten dargestellt, die jede für sich als eine Kompromisslösung auf der Skala zwischen Übersichtlichkeit und Genauigkeit zu verstehen ist.

2.1 Synoptische Darstellung (Unigramme, Top 10 Bigramme je Buchstabe, Top 100 aller Bigramme)

In der synoptischen Darstellung werden im ersten Abschnitt alle Buchstaben (Unigramme) mit ihrer Unigramm-Häufigkeit absteigend sortiert präsentiert entsprechend der ersten Spalte in Tab. 4. Hierbei sind alle 53 Basiszeichen berücksichtigt, die Häufigkeit der Umlaute ist der ihrer Basiszeichen zugeschlagen.

# häufigste Buchstaben	# häufigste Nachfolger von A	# häufigste Buchstabenkombinationen
150.000 e 96.000 n 81.000 i […]	1.100 Au 660 Ab 640 Al […] […]	34.000 ch 33.000 en 29.000 er […]
alle 53 Basiszeichen (Umlaute gemappt)	*pro berücksichtigte Buchstaben je max. 10 (ggf. weniger, wenn keine 10 belegt)*	*Top 100*

Tab. 4 Abschnittsweise Angaben in der synoptischen Darstellung

Im mittleren Abschnitt werden für alle drei Kategorien (Nachfolgebuchstaben von Großbuchstabe, Nachfolgebuchstaben von Kleinbuchstabe, Vorgängerbuchstaben von Kleinbuchstabe) blockweise je Buchstabe untereinander die zehn häufigsten Bigramme zweispaltig mit ihren *ipm*-Werten angegeben (entsprechend der zweiten Spalte in Tab. 4).

Im letzten Abschnitt werden entsprechend der dritten Spalte in Tab. 4 die 100 häufigsten Bigramme aus der nach Häufigkeit absteigend sortierten Gesamtrangliste angegeben. Diese Auflistung wird in drei Varianten angeboten: bezogen auf alle Bigramme, bezogen auf die

Bigramme, die mit einem Großbuchstaben beginnen, sowie bezogen auf die Bigramme, die mit einem Kleinbuchstaben beginnen.

Diese Angaben werden in einer utf8-kodierten txt-Datei bereitgestellt, alle *ipm*-Werte sind auf Stelligkeit 2 gerundet.

2.2 Visuelle Darstellung der Bigrammverteilung

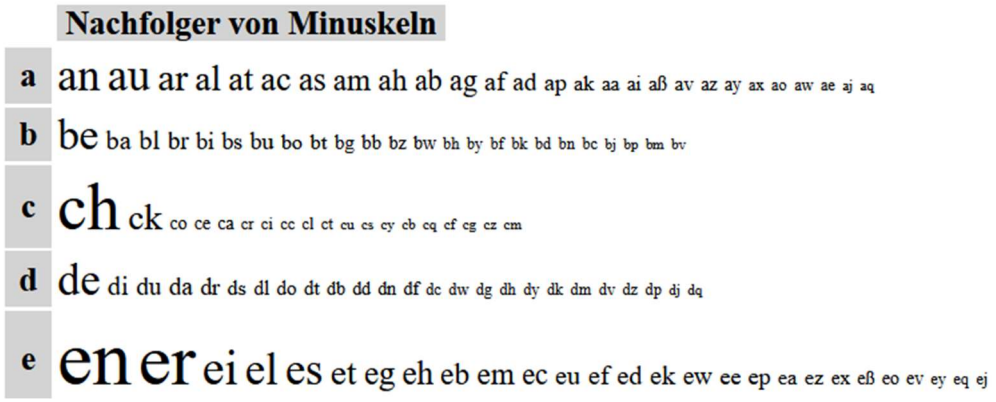

Abb. 2 Visuell unterstützte Präsentation aller Bigramme eines Bezugsbuchstabens (Ausschnitt a-e , nach rechts absteigend sortiert)

Für alle drei Kategorien (Nachfolgebuchstaben von Großbuchstabe, Nachfolgebuchstaben von Kleinbuchstabe, Vorgängerbuchstaben von Kleinbuchstabe) werden je Bezugsbuchstabe zeilenweise die ermittelten Bigramme aufgeführt. Beginnend mit dem häufigsten Bigramm folgen alle weiteren absteigend sortiert; die Schriftgröße drückt dabei entsprechend Häufigkeitsstufen aus basierend auf den auf Stelligkeit 2 gerundeten *ipm*-Werten, die hier nicht angezeigt werden.

Diese Darstellungsform ist html-/css-basiert als Webseite aufrufbar.

2.3 Gesamtdarstellung der Bi- und Unigrammhäufigkeiten

char,ipm A,3889 B,2783 C,1067 […]	char1,a,**b**,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,ß,t,u,v,w,x,y,z **A**,2,**656**,75,73,0,11,10,77,2,2,18,644,92,641,1,42,2,335,45,,60,1073,10,0,3,0,0 B,347,0,,,836,,,,338,0,,471,,,235,,,327,,,,223,,,,0, C,157,,,,40,,,146,43,,,43,,,587,,,41,,,,2,,,,3, […]
.UG.csv	.BG.csv

Tab. 5 Gesamttabellen der ganzzahlig gerundeten ipm-Werte der Uni- und Bigramme

Alle Ergebnisse der Uni- und Bigrammzählung sind ganzzahlig gerundet in zwei Tabellen zusammengefasst, die als csv-Dateien kodiert abgespeichert sind. In der Form sind speziell die Bigramm-Angaben für den Menschen unübersichtlich und schwer handhabbar. Sie lassen sich

aber in ein Tabellenkalkulationsprogramm wie MS Excel oder Libre/Open Calc einlesen[6] und dann mit den dort angebotenen Möglichkeiten übersichtlicher gestalten und auswerten.

char1	a	**b**	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r
A	2	**656**	75	73	0	11	10	77	2	2	18	644	92	641	1	42	2	335
B	347	0			836				338	0		471			235			327
C	157				40			146	43			43			587			41

Tab. 6 Ausschnitt der obigen, nach LibreCalc importierten Bigramm-csv-Datei

Die erste Zeile beinhaltet die Spaltenüberschriften. Entsprechend steht in den folgenden Zeilen im Feld der ersten Spalte der erste Buchstabe des Bigramms (also untereinander alle großen und kleinen Buchstaben). Die weiteren Spalten stehen für den zweiten Buchstaben des Bigramms (nebeneinander nur die kleinen Buchstaben). Eine Fehlanzeige, d.h. ein nicht beobachteter Bigrammtyp, wird in der csv-Datei als nicht angegebener Wert zwischen zwei Kommata kodiert; die Tabellenkalkulationsprogramme setzen dies üblicherweise als leeres Feld um. Ein *ipm*-Wert von 0 ist das Ergebnis einer Zahl zwischen $0 < x < 0{,}5$. Alle Werte sind nicht auf die sonst verwendete Stelligkeit 2 vereinfachte, sondern lediglich ganzzahlig gerundete Werte.

Der Wert im Feld der Zeile für den Buchstaben ‚A' in der Spalte für den Buchstaben ‚b' ist z.B. der ganzzahlig gerundete *ipm*-Wert des Bigramms ‚Ab': 656. Dieser Wert wurde für die obigen Darstellungen auf Stelligkeit 2 gerundet (660) verwendet für (1) Häufigkeit des ‚b' als Nachfolgebuchstabe des ‚A', sowie (2) Häufigkeit des ‚A' als Vorgängerbuchstabe des ‚b'.

3. Vergleichswerte einer breiteren Datengrundlage

Um die Ergebnisse dieser Studie vor dem Hintergrund einer breiteren Datengrundlage einordnen zu können, wurden alle Berechnungen auf einer aktuellen Fassung desselben Korpus durchgeführt, das den DeReChar-Auswertungen (Institut für Deutsche Sprache, 2018) zugrundegelegt wurde (im Folgenden DRC genannt).

		Datengrundlage	
		DeReChar-Korpus DRC (2018/2021)	KJL
Tokenbegriff	alle/ausgewählte Zeichen eines Tokens eines „typisch" deutschen Textes	(s. Institut für Deutsche Sprache, 2018)	
	Basiszeichen des deutschen Alphabets nur ausgewählter Token (s.o.)	.DRC.-Dateien	.kjl.-Dateien

Tab. 7 Einordnung der Vergleichsdaten der breiteren Datengrundlage

[6] als Feldtrenner bitte ausschließlich das Komma auswählen, als Kodierung sollte jeder Zeichensatz funktionieren, der ASCII-/Latin-/ISO-8859/UTF-8 basiert ist, da keine nicht-Latin Zeichen enthalten sind

Da der gleiche Tokenbegriff wie bei den KJL-Auswertungen verwendet wurde, können die Effekte durch dessen Einschränkungen geprüft werden. Für Aussagen anderer und allgemeinerer Art sind diese Daten nur bedingt geeignet, sie müssen auf jeden Fall mit Blick auf den Tokenbegriff interpretiert und ggf. relativiert werden.

In der aktuellen Fassung des DRC-Korpus wurden 1.431 verschiedene Bigramm-Typen mit weit über 108.000.000.000 Instanzen ermittelt bei einer durchschnittlichen Wortformlänge von 5,89 Unigrammtoken pro Wortformtoken (vgl. Tab. 8).

	Anzahl Token	**Anzahl Types**	**Anzahl ignorierte Token (Types)**
Wortformen	22.086.076.399	29.380.110	3.822.439.692 (56.275.459)
Unigramme	130.149.641.955	53	11.771.368.060
Bigramme	108.063.565.556	1.431	

Tab. 8 Zusammenfassung der ausgewerteten bzw. ignorierten Größen im DRC-Korpus

Die Ergebnisse dieser Vergleichsauswertung werden auf dieselbe Art dargestellt wie die der KJL-Untersuchungen. In der synoptischen Darstellung werden jedoch aufgrund des größeren Korpusumfangs alle *ipm*-Werte auf Stelligkeit 3 gerundet.

4. Quellenverzeichnis

Institut für Deutsche Sprache (2018): *DeReChar v-uni-XXX-2018-02-28-1.0*, Korpusbasierte Zeichenhäufigkeitslisten zum Deutschen Referenzkorpus DeReKo, http://www.ids-mannheim.de/derewo, Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland. (verfügbar unter http://www.ids-mannheim.de/fileadmin/kl/derewo/derechar-v-uni-XXX-2018-02-28-1.0.pdf, letzter Zugriff am 31.10.2021)

Leibniz-Institut für Deutsche Sprache (2021): *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2021-I* (Release vom 02.02.2021). Mannheim: Leibniz-Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo

Steinig, Wolfgang (2017): *Grundschulkulturen. Pädagogik – Didaktik – Politik.* Berlin: Erich Schmidt.

5. Kontakt

Falls Sie speziellere Informationen benötigen, als dieses Werk bereithält, wenden Sie sich bitte an derewo@ids-mannheim.de.

Bei Veröffentlichung auf diesem Werk aufbauender Forschungsergebnisse bitten wir um eine kollegiale Mitteilung an derewo@ids-mannheim.de.

6. Lizenzbestimmungen

(zu zitieren als:)

DERECHAR-v-[bi|uni]-[KJL|DRC]-2021-10-31-1.0, Korpusbasierte Zeichenbi- und –uni-grammhäufigkeitslisten zum Kinder- und Jugendliteraturkorpus des Deutschen Referenzkorpus

DEREKO, http://www.ids-mannheim.de/derewo, © Leibniz-Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2021.

Die Bi-/Unigrammhäufigkeitslisten und die Dokumentation bilden eine Einheit. Diese Lizenzbestimmung darf aus keinem der Dokumente entfernt werden.

Dieses Werk ist unter die Creative Commons-Lizenz (by) gestellt (http://creativecommons.org/licenses/by/4.0/deed.de).

Namensnennung 4.0 International (CC BY 4.0)

Dies ist eine allgemeinverständliche Zusammenfassung der Lizenz (die diese nicht ersetzt).

Haftungsbeschränkung.

Sie dürfen:

- **Teilen** – das Material in jedwedem Format oder Medium vervielfältigen und weiterverbreiten
- **Bearbeiten** – das Material remixen, verändern und darauf aufbauen
- und zwar für beliebige Zwecke, sogar kommerziell.
- Der Lizenzgeber kann diese Freiheiten nicht widerrufen, solange Sie sich an die Lizenzbedingungen halten.

Unter folgenden Bedingungen:

- Namensnennung – Sie müssen angemessene Urheber- und Rechteangaben machen, einen Link zur Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Diese Angaben dürfen in jeder angemessenen Art und Weise gemacht werden, allerdings nicht so, dass der Eindruck entsteht, der Lizenzgeber unterstütze gerade Sie oder Ihre Nutzung besonders.
- Keine weiteren Einschränkungen – Sie dürfen keine zusätzlichen Klauseln oder technische Verfahren einsetzen, die anderen rechtlich irgendetwas untersagen, was die Lizenz erlaubt.

Hinweise:

- Sie müssen sich nicht an diese Lizenz halten hinsichtlich solcher Teile des Materials, die gemeinfrei sind, oder soweit Ihre Nutzungshandlungen durch Ausnahmen und Schranken des Urheberrechts gedeckt sind.
- Es werden keine Garantien gegeben und auch keine Gewähr geleistet. Die Lizenz verschafft Ihnen möglicherweise nicht alle Erlaubnisse, die Sie für die jeweilige Nutzung brauchen. Es können beispielsweise andere Rechte wie Persönlichkeits- und Datenschutzrechte zu beachten sein, die Ihre Nutzung des Materials entsprechend beschränken.