

Korpuszusammensetzung und Verlässlichkeit des deutschsprachigen Google Ngram-Viewers

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Als Google im Jahre 2009 die erste Version des Ngram-Viewer publizierte, hat die Digital Humanities-Community recht schnell die positiven und negativen Aspekte dieses Werkzeugs analysiert: Die Möglichkeiten einer wort- und begriffsgeschichtlichen Forschung sind sprunghaft erweitert worden, wie auch der begleitende Aufsatz von Michel deutlich belegte (Michel et al. 2011). Dessen teilweise zu naiven Umgang mit dem Quellenmaterial machte aber auch deutlich, dass die Autorinnen und Autoren kein Bewusstsein für die möglichen Fällen von korpusbasierten Forschungen hatten. Die wechselnde Zusammensetzung der zugrundeliegenden Textsammlung, die Unmöglichkeit auf die dahinterliegenden Texte zuzugreifen, das Fehlen von Metadaten für die Texte, falsche Jahreszahlen, OCR-Fehler u.a.m. sind dem Ngramm-Korpus wiederholt vorgeworfen worden (z.B. Underwood 2012). Die Arbeit von (Pechenick et al. 2015) hat gezeigt, wie die zunehmende Menge von wissenschaftlicher Literatur die Korpusanteile in der zweiten Hälfte des 20. Jahrhundert merklich verschiebt; unklar bleibt allerdings, ob dies nicht auch eine gesellschaftliche Entwicklung reflektiert, also keineswegs nur als Manko zu betrachten ist. Besonders einschlägig ist die Arbeit (Koplenig 2017), in der Veränderungen der Korpuszusammensetzung während des zweiten Weltkriegs untersucht werden: „the German GB corpus was strongly biased toward volumes published in Switzerland during WWII“ (Koplenig 2017)

Google hat zwei größere Updates vorgelegt, die die zugrundeliegende Textmenge deutlich erweitert haben. Hier die Entwicklung des Umfangs der deutschsprachigen Korpora, auf die ich mich im Folgenden beschränke:

	Tokens	Bücher
2009	37.439.210.527	406.666
2012	64.784.628.286	657.991
2019	286.463.423.242	3.843.962

Dadurch dass die Anzahl der digitalisierten Bücher noch einmal deutlich gesteigert werden konnte – und das über den gesamten Zeitraum, den der Viewer abdeckt, – scheinen die Fragen nach einem Bias der Auswahl weniger relevant zu werden. Entsprechend wurde und wird der Ngram-Viewer auch weiterhin in vielen Kontexten als

schnell zugängliches Werkzeug verwendet, das die Möglichkeiten einer auf sehr großen Datenbeständen basierenden Forschung anschaulich macht (z.B. Chen and Yan 2016, Gonçalves et al. 2018, Richey and Taylor 2020). Insgesamt gehören der *Ngram-Viewer* und die zugrundeliegenden freiverfügbaren Daten trotz aller berechtigten Kritik für viele Forschende zu den wichtigsten Daten-Publikationen der letzten Jahrzehnte.

In diesem Sinne wollte auch eine Kollegin, die zu Fragen der literarischen Kanonisierung arbeitet, den Viewer verwenden, aber bei der Analyse der Ergebnisse fiel uns schnell auf, dass die Namen einer Reihe der hochkanonischen deutschsprachigen Autoren – Thomas Mann, Goethe, Schiller, Kafka, Brecht – nach 2005 einen auffälligen Abwärtstrend aufweisen (siehe Fig. 1).

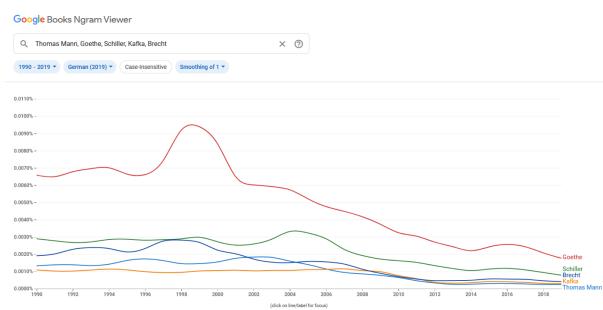


Fig. 1: Kanonisierte Autoren (Goethe, Schiller, Thomas Mann, Kafka, Brecht) 1990-2019.

Das Phänomen zeigt sich noch deutlicher, wenn man sich nur für kanonisierte Autor *innen* interessiert (siehe Fig. 2).

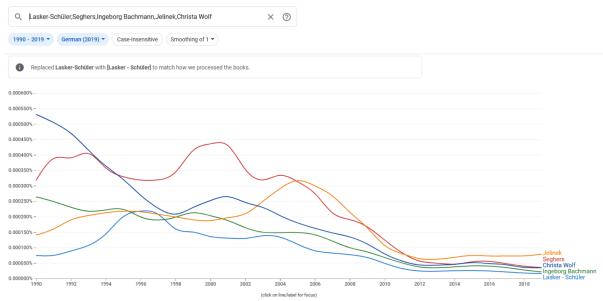


Fig. 2: Häufigkeit kanonisierter Autorinnen im Ngram-Viewer (Lasker-Schüler, Seghers, Bachmann, Jelinek, Christa Wolf) 1990-2019.

Bevor wir nun allgemeinere kulturanalytische Thesen über das Ende der Bildungskultur formulierten, wollte ich die Solidität der Daten prüfen. Doch wie kann man ein Korpus auf möglichen Bias untersuchen, wenn weder die Liste der Texte geschweige die Texte selbst vorliegen, sondern nur eine Reihe von generischen Metadaten und 1-5 Gramme?

Die Korpusanomalien

Da die NGramme, die dem *Ngram-Viewer* zugrundeliegen, ebenfalls publiziert sind, ist es naheliegend, erst ein-

mal die Rohwerte und deren Entwicklungstendenzen zu untersuchen. Das Ergebnis (Fig. 3) zeigt zumindest für drei der untersuchten Namen eine einheitliche Tendenz: aufsteigend – also genau das Gegenteil der absteigenden Entwicklung, die der *Ngram-Viewer* präsentiert. (Auffällig ist außerdem ein Abfallen in allen Verteilungen im Jahre 2010.)

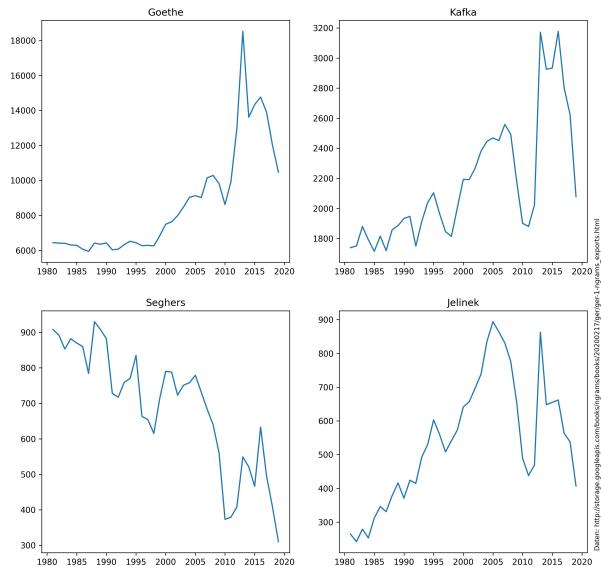


Fig. 3: Rohwerte für die Häufigkeiten im NGram-Viewer Korpus (1-Gramme, 2019)

Wie lässt sich der Widerspruch zwischen den Ergebnissen erklären? Der *Ngram-Viewer* zeigt die *relativen* Häufigkeiten der Wörter an, d.h. den Anteil den das Wort, z.B. der Name ‚Goethe‘, an der Menge aller publizierten Wörter dieses Jahres hat. Dadurch kann man Frequenzen aus dem 18. Jahrhundert, die nur auf einigen wenigen Büchern beruhen, mit denen im 21. Jahrhundert vergleichen. Die Differenz der Ergebnisse könnte also so erklärt werden, dass zwar die Anzahl an Nennungen von Goethe weiter ansteigt, aber ab ca. 2005-2010 zugleich sehr viel mehr Bücher im Korpus sind, in denen die Namen der Autoren und Autorinnen nicht genannt werden. Das würde bedeuten, dass insgesamt im Korpus die Anzahl der Bücher pro Jahr deutlich angestiegen sein müsste. Die entsprechenden Daten sind im Ngram-Korpus vorhanden und sie bestätigen die These:

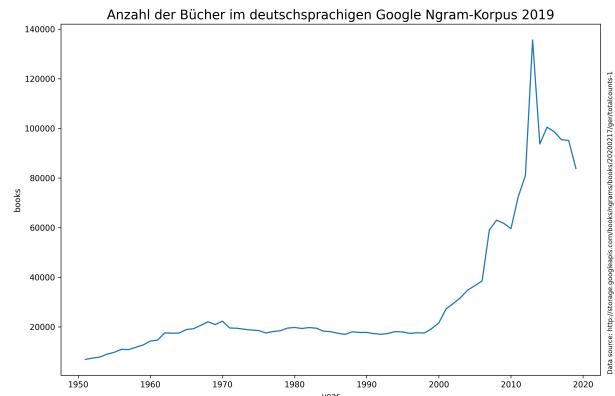


Fig. 4: Anzahl der Bücher pro Jahr im DE-Ngramm-Korpus 2019

Wenn die Rohdaten aus Fig. 3 nun mit den Daten über die Anzahl der Token pro Jahr normalisiert werden, dann ergibt sich der Trend, der im Ngram-Viewer sichtbar wurde, alle Werte stürzen nach 2005 mehr oder weniger steil ab.

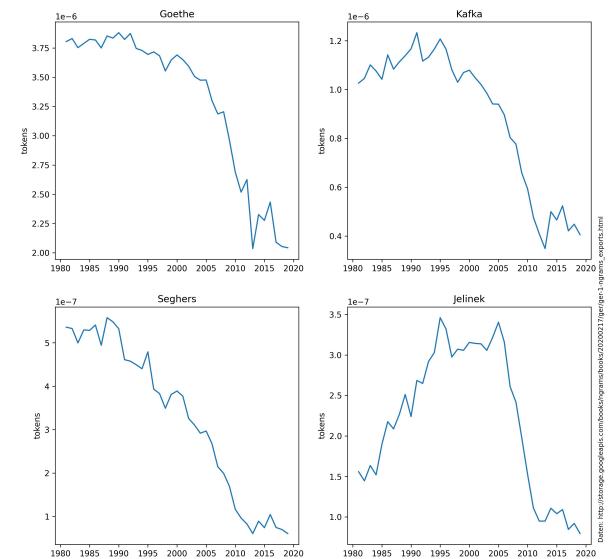


Fig. 5: Häufigkeitswerte der kanonisierten Autoren und Autorinnen geteilt durch die Anzahl der Token pro Jahr

Allerdings ist der sehr steile Anstieg der Buchzahlen in Fig. 4 ziemlich überraschend. Wir wissen, dass Google Books auf den Ergebnissen der Digitalisierungskampagne Googles in Kooperation mit einer ganzen Reihe von internationalen Forschungsbibliotheken beruht. Die deutschsprachigen Ergebnisse verdanken sich nicht zuletzt den Kooperationen mit der Österreichischen Nationalbibliothek und der Bayerischen Staatsbibliothek. Deinen Bestände speisen sich in den letzten Jahrzehnten aus Pflichtabgabeexemplaren und einer umfassenden Erwerbspolitik. Woher kommt also der plötzliche Anstieg? Sind – den Klagen der Verlage zum Trotz – seit 2005 sehr viel mehr Bücher als früher gedruckt worden?

Für die Buchproduktion konnte ich zwei Quellen verwenden, die die Daten gleich in digitaler Form anbie-

ten: Statista, ein kommerzieller Datenanbieter, der Zahlen des Börsenvereins des deutschen Buchhandels zur Anzahl der Neuerscheinungen aufbereitet hat (Börsenverein 2022),¹ und die Angaben über die Anzahl der Buchpublikationen insgesamt in Thomas Rahlf's Zeitreihen zur historischen Statistik (Rahlf 2015). Sie decken allerdings nicht die gleichen Zeiträume ab. Statista hat die aktuelleren Daten, reicht aber nicht soweit zurück, während Rahlf's Datenreihe schon 2015 endet. Schon ein erster Blick auf die Fig. 6, die die Anzahl Bücher, die laut Börsenverein/Statista zwischen 2002 und 2019 gedruckt worden sind, mit der Anzahl der Bücher vergleicht, die dem Ngramm-Korpus zugrundeliegen, zeigt etwas Verblüffendes: In dem Korpus sind nach 2012 mehr Bücher enthalten, als Neuerscheinungen in dem Jahr gedruckt wurden.

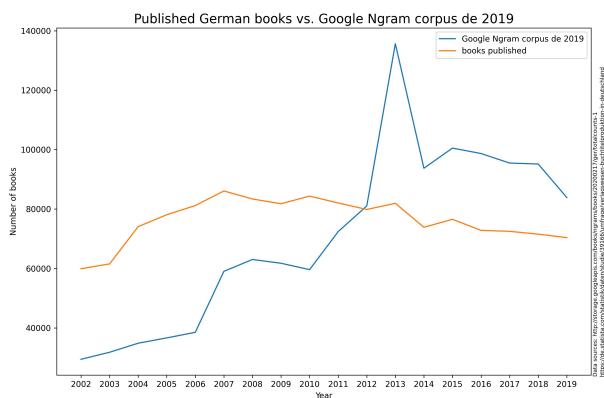


Fig. 6: Bücher im deutschsprachigen Ngramm-Korpus und die Anzahl der jährlich publizierten Bücher (Statista)

Vielleicht stimmen ja die Zahlen von Statista nicht. Vielleicht sind die Angaben in dem Ngram-Korpus aus dem Jahre 2019 falsch. In der Fig. 7 sind nun alle Informationen gemeinsam eingeflossen. Sie zeigt zum einen die Werte der Buchproduktion aufgrund der Daten des Börsenvereins – das ist der kurze violette Graph – und nach den Werten von Rahlf – die rote Kurve ab 1950. Es ist offensichtlich, dass die Werte zur Buchproduktion sich voneinander unterscheiden: Rahlf liegt immer etwas höher, da er ja nicht nur die Neuerscheinungen erfasst, aber beide beschreiben ziemlich parallel die gleiche Dynamik. Bis ca. 2009 steigt die Produktion immer weiter an und fällt danach ab. Aber welche der beiden Kurven man auch zugrundelegt, stets übersteigen die Zahlen des Ngramm-Korpus die Werte im Jahr 2013 und wohl auch danach.

Die Grafik enthält neben den Werten für das aktuelle Korpus von 2019 auch die Werte für die früheren Ngramm-Korpora von Google aus den Jahren 2009 und 2012, die deutlich kleiner waren. Beginnen wir bei den Werten vor 1995: Es ist auffällig, dass Google mit dem letzten Update den Anteil an der Buchproduktion eines Jahres, der digitalisiert vorliegt, deutlich steigern konnte, so dass bis in die Mitte der 1960er Jahre teils 50% und mehr im Korpus enthalten sind. Für den Zeitraum von den späten 1960ern bis in die späten 1990er stagnieren die Werte der Korpora, während die Buchproduktion in diesen Jahren steil angestiegen ist. Während im Jahr 1967 erstaunliche 67% der nach Rahlf's publizierten Bücher im

Korpus liegen, sind es 1997 „nur“ noch 23%. Erst danach, 1998 bis 2006, wächst der Anteil wieder. In den Korpora von 2009 und 2012 geschieht dies noch relativ langsam, während die Erweiterung von 2019 hier deutlich stärker zulegt. Von 2006 auf 2007 springen die Werte allerdings steil nach oben. Das gilt für alle drei Stufen des Korpus, aber auch hier ist der Anstieg im 2019-Korpus noch einmal deutlich ausgeprägter. Danach, zwischen 2008 und 2010 verhalten sich die Werte im Korpus auf einem hohen Niveau parallel zu den Werten der Buchproduktion und fallen mit dieser sogar leicht ab. Das ändert sich wiederum 2011-2013, wo wir einen weiteren Anstieg beobachten können, der diesmal sogar das Niveau der Buchproduktion übertrifft.

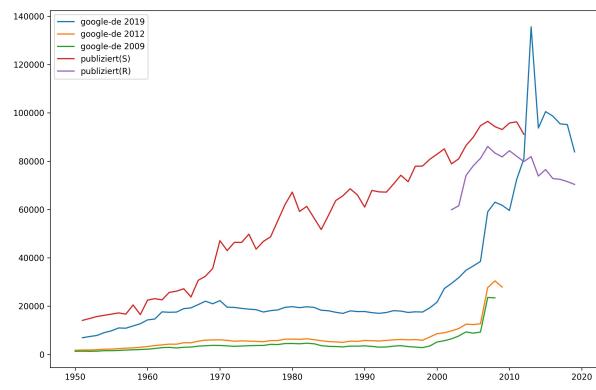


Fig. 7: Entwicklung der Buchproduktion und der Google Korpora 1950-2019

Wenn wir noch einmal auf Grafik 3 und 4 blicken, dann sehen wir dort, dass die Rohwerte für Goethe und zwei andere kanonisierte Autoren in den späten 1990er Jahren deutlich ansteigen, sie zugleich in der normalisierten Darstellung schon abfallen. Das bedeutet, dass bereits in der Phase von 1998-2006 „andere“ Texte hinzugefügt wurden, deren Zusammensetzung anders war als das bisherige Korpus. Um welche Texte könnte es sich hierbei handeln?

Faktoren der Korpusverzerrung nach 1995

Die oben erwähnten verschiedenen Phasen der Veränderung legen es nahe zu vermuten, dass nicht ein einzelner Eingriff in das Korpus, sondern eine Reihe verschiedener Texthinzufügungen die beobachteten Phänomene bedingen. In einem anderen Projekt, in dem es um die Analyse von Hefromanen geht, war bereits aufgefallen, dass sich die einschlägigen Verlage der Komplexität der Feststellung der richtigen Metadaten, insbesondere des Publikationsdatums, dadurch entledigt haben, dass sie einfach alle retrodigitalisierten Texte unter dem Datum veröffentlichen, an dem die digitale Kopie publiziert wird. Um zu testen, ob diese Texte auch im Ngram-Korpus sind, wurden die entsprechenden Serienautoren und -helden gesucht:

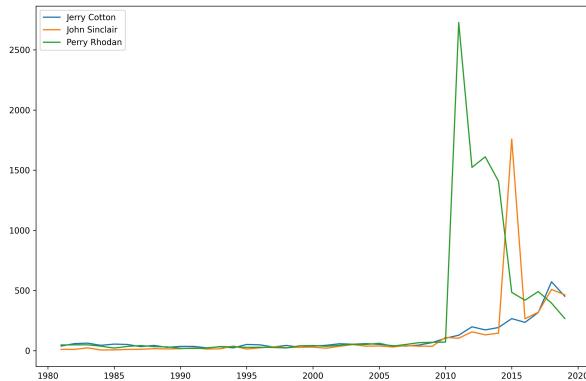


Fig. 8: Häufigkeit von Heftroman-Helden im Ngramm-Korpus (2019)

Die deutliche Steigerung ab 2010 spricht dafür, dass auch hier die Retrodigitalisate unter dem jeweiligen Jahr aufgeführt sind. Allerdings können selbst einige Tausend Heftromane nicht alleine verantwortlich sein.

Wie könnte man nun weitere Faktoren erkennen? Ein offensichtlicher Weg geht über das sprachliche Material. Die neuen Texte würden für bestimmte Worte zu einer relativen Erhöhung führen, selbst wenn viele andere Worte einen relativen Rückgang aufweisen. Da Google die Daten im 2019-Korpus mit Wortklasseninformation ausliefer, war es einfach, alle Substantiv aus den 1-Grammen zu extrahieren, rd. 13 mio. Anschließend wurden die Substantive herausgefiltert, die von 1995 bis 2019 jedes Jahr auftauchen und zwar insgesamt mindestens 2 500 mal. Für diese restlichen rd. 350.000 Wörtern wurde für die Daten jedes Wortes eine lineare Regression berechnet und die Steigung der Geraden als Filterungsfaktor verwendet, um die rd. 250 Wörter zu finden, die in dieser Zeit den größten Anstieg verzeichnen.

Eine manuelle Sichtung dieser Wortliste zeigte den Einfluss mehrerer Textsorten. Vor allem aber fiel der einzige Name in der Liste auf: GRIN. Es handelt sich um eine deutsche Verlagsgruppe, zu der neben dem GRIN-Verlag selbst u.a. auch die Webseite hausarbeiten.de gehört. Der Verlag, der von Beobachtern als ‚vanity publisher‘ oder ‚predatory publisher‘ (Shrestha 2021) eingeschätzt wird, publiziert alle Texte digital oder als Book on Demand. Eine „Lektorierung findet nicht statt“ (Wikipedia). Laut Verlagswebseite wurden bis ins Jahr 2018 200.000 Texte publiziert. Seitdem sind schätzungsweise mindestens weitere 40.000 Titel publiziert worden.² Fig. 9 zeigt die Anzahl der Bücher im NGramm-Korpus, in denen das Token ‚GRIN‘ vorkommt, was wohl weitgehend identisch ist mit Titeln des Verlags.

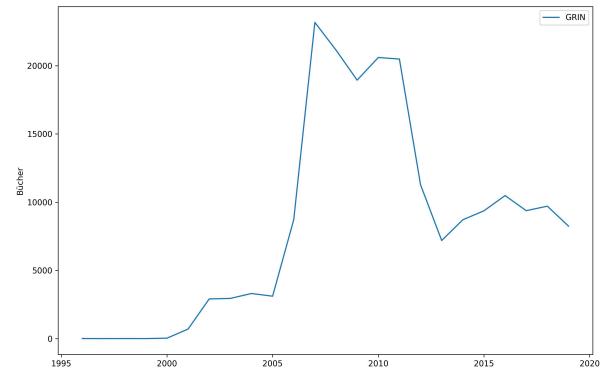


Fig. 9: Anzahl der Bücher im Ngramm-Korpus (2019), in denen das Wort 'GRIN' vorkommt.

Die Texte des GRIN Verlags haben zwar alle eine ISBN-Nummer, aber die meisten sind deutlich kürzer als Bücher im herkömmlichen Sinn, viele sind Aufsätze, die nur 20-30 Seiten lang sind. Beide Faktoren, die Ummeiste an quasi-wissenschaftlichen kurzen Texten des GRIN-Verlags und die falsch datierten Heftromane führen dazu, dass die durchschnittliche Länge von Büchern im Ngramm-Korpus sich seit 2000 deutlich verringert hat (siehe Fig. 10), allerdings sind die Werte seit 2017 fast wieder auf dem alten Niveau:



Fig. 10: Durchschnittliche Anzahl der Seiten pro Buch im Ngramm-Korpus

Fassen wir zusammen: Nach 1998 ändert sich die Zusammensetzung des deutschsprachigen Ngramm-Korpus einschneidend, so dass es für die meisten Analysen zur Entwicklung von Sprache und Kultur weitgehend unbrauchbar wird. Dazu tragen eine Reihe von Faktoren bei, von denen zwei identifiziert werden konnten: Schwerwiegender ist, schon aus Umfanggründen, der Anteil der Publikationen des GRIN-Verlags. Sie geben zwar einen Einblick in eine bestimmte Form universitärer Wissenschaftskommunikation, haben aber nichts mit der sonstigen Buchproduktion zu tun. Hinzukommen die falsch datierten Retrodigitalisierungen einiger Verlage. Zugleich zeigt die Analyse der Daten, dass dies nicht die einzigen Faktoren sind, die hier ins Gewicht fallen. Wenn man sich auf die Wörter mit den steilsten Karrie-

ren in den letzten 25 Jahren konzentriert (Fig. 11), dann fällt auf, dass dies allgemeine Token sind, die sich eher in Romanen als in Fachtexten finden (besonders die Anführungszeichen, mit denen in den meisten deutschen Drucktexten direkte Rede markiert wird): „Augen Blick Du Frau Gesicht Hand Kopf Leben Mal Mann Moment Mutter Stimme Tag Tür Vater « »“ Da zugleich die Länge ansteigt und die Anzahl der Texte sehr hoch bleibt, außerdem diese Texte aber wohl nicht in den offiziellen Verlagsstatistiken auftauchen, handelt es sich vermutlich um Texte aus literarischen *Selfpublishing* Verlagen, die in der Umfrage des Börsenvereins nicht miteinbezogen waren. Darüber, ob diese nicht doch Teil eines Kulturgrafen sein sollten, lässt sich allerdings trefflich streiten.

An diese explorative Studie könnte nun eine Untersuchung anschließen, die die Veränderungen der Korpuszusammensetzung als überdurchschnittlich starke Veränderung der Token-Verteilungen formalisiert (Koplenig 2017) und so den hier etwas vernachlässigten Aspekt, wann sich genau die Veränderungen ergeben, herausarbeitet. Die ‚typische‘ Verwendung des Ngramm-Korpus und -viewers, nämlich die Untersuchung der Verwendungshäufigkeit von Termen in der schriftlichen Öffentlichkeit, ist durch die starken Schwankungen in der Zusammensetzung des Korpus sehr fragwürdig geworden. Da nach 2000 sonst eher randständige Bereiche, quasi-wissenschaftliche Texte und die Produktion der selbstverlegten Autorinnen und Autoren, das Korpus dominieren, ist es auch für rein sprachanalytische Untersuchungen, etwa zur Kollokationsanalyse, kaum verwendbar. Aber insgesamt verdient die Frage, ob und unter welchen Vorzeichen die Daten nicht doch für bestimmte Analysen herangezogen werden können, eine genauere Untersuchung.

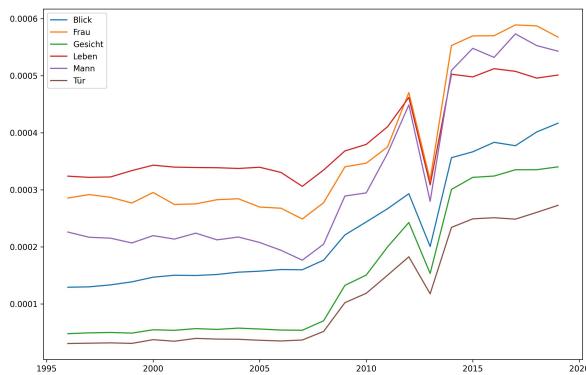


Fig. 11: Relative Häufigkeiten einiger Wörter, die 1995 bis 2019 zunehmend häufiger verwendet werden.

Bibliographie

- Börsenverein des deutschen Buchhandels** (2022) *Buchtitelproduktion: Anzahl der Neuerscheinungen in Deutschland in den Jahren 2002-2021*. Statista. Available at: <https://de.statista.com/statistik/daten/studie/39166/umfrage/verlagswesen-buchtitelproduktion-in-deutschland/> (Accessed: 8 March 2022).
- Chen, Y. and Yan, F.** (2016) ‘Centuries of Sociology in Millions of Books’. Available at: <https://doi.org/10.1111/1467-954X.12399>.
- Gonçalves, B. et al.** (2018) ‘Mapping the Americanization of English in Space and Time’, *PLOS ONE*, 13(5), p. e0197741. Available at: <https://doi.org/10.1371/journal.pone.0197741>.
- Koplenig, A.** (2017) ‘The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII’, *Digital Scholarship in the Humanities*, 32(1), pp. 169–188. Available at: <https://doi.org/10.1093/llc/fqv037>.
- Michel, J.-B. et al.** (2011) ‘Quantitative Analysis of Culture Using Millions of Digitized Books’, *Science*, 331(6014), pp. 176–182. Available at: <https://doi.org/10.1126/science.1199644>.
- Pechenick, E.A., Danforth, C.M. and Dodds, P.S.** (2015) ‘Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution’, *PLOS ONE*, 10(10), p. e0137041. Available at: <https://doi.org/10.1371/journal.pone.0137041>.
- Rahlf, T.** (2015) *Deutschland in Daten: Zeitreihen zur Historischen Statistik*. Bonn: Bundeszentrale für politische Bildung.
- Shrestha, J.** (2021) *Vanity publishers: How to identify and avoid them*. Available at: <http://eprints.rclis.org/42635/> (Accessed: 2 August 2022).
- Underwood, T.** (2012) ‘How not to do things with words’, *The Stone and the Shell*, 25 August. Available at: <https://tedunderwood.com/category/ngrams/> (Accessed: 1 August 2022).

Fußnoten

- Der Börsenverein dient in erster Linie der Interessenvertretung deutscher Firmen, auch wenn seit ca. 10 Jahren internationale Firmen Mitglied werden können.
- Die Deutsche Nationalbibliothek verzeichnetet 394.000 Treffer für den GRIN-Verlag, allerdings wurden mindestens seit 2017 der größere Teil der Titel doppelt verzeichnet. Die DNB listet für den Zeitraum von 2019 bis heute rd. 78.000 Titel des GRIN-Verlags.