Alexander Lasch

Maschinelle Stilanalyse im Literaturunterricht

Erstveröffentlichung in / First published in:

Der Deutschunterricht. 2019, (1), S. 87-96 [Zugriff am: 12.04.2022]. Friedrich Verlag. ISSN 0340-2258.

Final edited form was published in "Der Deutschunterricht". 2019, (1), S. 87-96. Friedrich Verlag. ISSN 0340-2258.

ALEXANDER LASCH

Maschinelle Stilanalyse im Literaturunterricht

1 Hinführung

Initiativen wie bspw. literaturlinguistik.de bemühen sich seit Langem darum, die Kluft zwischen Literatur- und Sprachwissenschaft zu überbrücken, indem man in übergreifenden Fragestellungen zeigt, dass die je verschiedenen Perspektiven auf den gemeinsamen Gegenstand sehr fruchtbar sein können. Relativ unverdächtig erscheinen in der aktuellen Diskussion die Konzepte des Autorenstils und damit die Frage nach Autorschaft; bzw. auf performativer und auf Textebene des Erzählers, wenn man sich auf eine basale Terminologie verständigt (z. B. die von Genette 2010) und einzelne Aspekte aus dem jeweiligen Fachbereich so betont, dass Analyseergebnisse im je anderen für Interpretationen fruchtbar gemacht werden können.[1] Das ist der eine Brückenschlag, den dieser Beitrag versucht. Der zweite betrifft den auch in den Schulen ausgetragenen Konflikt zwischen naturwissenschaftlichen und geisteswissenschaftlichen Fächern, der im Prozess der „Digitalisierung" auch in Bereiche dringt, die vorher als Domänen klar voneinander abgegrenzt waren.

Der Artikel wird ein Schlaglicht auf den Einsatz von korpuslinguistischen Methoden in der Literaturanalyse setzen (Kap. 2.2) und einen sehr knappen Forschungseinblick in die Forensische Linguistik bieten (Kap. 2.3 und 3). Der Schwerpunkt (Kap. 4) allerdings wird auf der Beschreibung von Tools (und eines Workflows) liegen, die praktisch im Deutschunterricht verwendet werden können. Diese werden am Beispiel der Hypothesengenerierung für die Interpretation von Goethes *Faust*-Dramen vorgestellt.

2 Forschungsschlaglichter

2.1 Korpuslinguistische Methoden für die Literaturanalyse

Die deutschsprachige Forschungsgemeinschaft nimmt das Thema des Einsatzes maschineller Analysen für die Untersuchung literarischer Texte nur sehr zaghaft auf (vgl. aber die wenigen Beispiele wie etwa Herrmann/Lauer 2018); die angloamerikanische Forschung ist hier deutlich progressiver. Zu empfehlen ist für einen ersten Überblick der Besuch der Online-Präsenz des „Stanford Literary Lab" (https://litlab.stanford.edu/, 8.6.2018), das in regelmäßigen Abständen „Literary Lab Pamphlets" als „Open Access"-Publikationen zur Verfügung stellt und über neue Entwicklungen am Schnittpunkt zwischen maschineller und Analyse literarischer Texte informiert. Herauszuheben ist Franco Morettis (2017) „Patterns and Interpretation", der die Folgen der ‚neuen Skalierung' im Umgang mit dem Gegenstand betont:

> One thing for sure: digitization has completely changed the literary archive. People like me used to work on a few hundred nineteenth-century novels; today, we work on thousands of them; tomorrow, hundreds of thousands. This has had a major effect on literary history, obviously enough, […] but also on critical methodology; because, when we work on 200,000 novels instead of 200, we are not doing the same thing, 1,000 times bigger; we are doing a different thing. The new scale changes our relationship to our object, and in fact it changes the object itself.
>
> MORETTI 2017, 1

In diesen Dimensionen denkt vielleicht die deutschsprachige Linguistik, die Literaturwissenschaft aber noch keinesfalls. Vielleicht kann diese eher anschließen an Ergebnissen, die Irina Keshabyan und Ángela Almela 2012 in einem Sonderheft des *International Journal of English Studies* unter dem Titel *A New Approach to Literature: Corpus Linguistics* zusammenfassten. In kleineren Beiträgen werden vor allem die Arbeiten von Biber (2011), Fischer-Starcke (2010) und Mahlberg (2007a und 2007b) in den Mittelpunkt gerückt, die die Relevanz maschineller Patternanalysen für unterschiedliche Fragestellungen gezeigt haben.

Besonderen Einfluss auf das Forschungsfeld hat Fischer-Starcke (2010). Ihre Monographie zu *Jane Austen and her Contemporaries* ist musterbildend und hat zahlreiche weitere Publikationen zur Autorenstilistik angeregt. Michaela Mahl-

Final edited form was published in "Der Deutschunterricht". 2019, (1), S. 87-96. Friedrich Verlag. ISSN 0340-2258.

berg publiziert 2013 etwa *Corpus Stylistics and Dickens's Fiction*. Zu nennen ist weiter Balossi (2014), die sich mit literarischer Sprache und Charakterisierung in Virginia Woolfs *The Waves* auseinandersetzt. Doch auch einzelne Gattungen kommen nach der Auseinandersetzung mit Autorenstilen und -registern in den Blick: Hoover, Culpeper und O'Halloran (2014) wenden sich in *Digital Literary Studies* den klassischen literarischen Gattungen *Poetry, Prose and Drama* zu.

Heute ist das Thema in der angloamerikanischen Forschung längst Handbuchwissen. Im *Cambridge Handbook of English Corpus Linguistics* (CHECL), herausgegeben von Douglas Biber und Randi Reppen, publiziert Mahlberg 2015 den Artikel "Literary style and literary texts" selbstverständlich neben Artikeln zu Themen wie dialektaler Variation oder diachronen Registern.

1.2 Forensische Linguistik

Das Thema der Forensischen Linguistik ist seit Hannes Kniffkas frühem Artikel „Der Linguist als Gutachter" (1981) in der deutschsprachigen Forschungslandschaft präsent. Seine Arbeiten prägten über Jahrzehnte der linguistischen Auseinandersetzung mit der Autorschaftsidentifikation das Verhältnis zwischen Linguistik und Kriminalistik (vgl. exemplarisch Kniffka 1990, 1994, 2001 und 2007). Arbeiten wie die von Christa Dern zur *Autorenerkennung* (2009) oder der (einzige) einführende Band von Eilika Fobbe in die *Forensische Linguistik* (2011 und im *Handbuch Sprache im Recht* 2017) können heute exemplarisch für die deutsche Forschungslandschaft stehen: Sie beleuchten das Thema aus der Sicht der Kriminalistik bzw. der Linguistik und fokussieren dabei (meist) auf die Analyseeinheit „Text" und damit methodisch auf eine qualitative Textanalyse. Was die maschinelle Autorschaftsanalyse angeht, ist man in der Anwendung automatischer Verfahren zur Ermittlung von Autorschaft ("authorship identification") in der angloamerikanischen Forschung sehr viel weiter als in der deutschsprachigen Forschung. Das hat verschiedene Gründe: (1) Zum einen ist der Forensischen Linguistik als Bereich der Angewandten Linguistik in Deutschland kein eigener institutionalisierter Ausbildungs- oder Studiengang gewidmet. Damit sind die meisten Fachwissenschaftlerinnen und -wissenschaftler, die sich der Forensischen Linguistik zuwenden, in erster Linie für Fachbereiche der Phonetik, Textwissenschaft oder Korpuslinguistik ausgewiesen – es ist nach wie vor ein Nischenthema, wenn es auch in der interessierten Öffentlichkeit viel Aufmerksamkeit auf sich zieht. (2) Die meisten Fachwissenschaftlerinnen und Fachwissenschaftler, die sich der maschinellen forensischen Autorschaftsidentifikation widmen, arbeiten als Gutachterinnen und Gutachter in Ermittlungsverfahren eng mit Sicherheitsbehörden zusammen und publizieren in den seltensten Fällen ihre Ergebnisse.[2]

Deshalb wird der Ansatzpunkt hier sein, in die Grundlagen der forensischen Autorschaftsidentifikation einzuführen, in einem zweiten Schritt kurz die maschinellen Analysemöglichkeiten zu reflektieren (Kap. 3) und in einem dritten Schritt auf den Gegenstand literarischer Texte anzuwenden (Kap. 4).

3 Multifaktorielle Stilanalyse und Autorschaftsidentifikation in der Forensischen Linguistik

Ich möchte beginnen mit einigen Prämissen der Forensischen Linguistik das Autorschaftskonzept betreffend. Das muss ich tun, um die theoretischen und methodischen Grundlagen für meinen heutigen Gehversuch zur Frage von Autorschaft zu wagen. Ich hoffe zu zeigen, dass im Hinblick auf Text- bzw. Autorschaftsstile die unterschiedlichen Autorschaftskonzepte nicht willkürlich und unreflektiert nebeneinandergestellt werden können.

Forensische Linguistik ist ein „Teilbereich der Linguistik, der die Analyse solcher sprachlichen Daten (einschließlich ihrer Präsentation vor Gericht) umfasst, die Gegenstand juristischer Betrachtung sind." (Fobbe 2011, 16) Dazu zählen bereits nach Levi (1982) und dann auch Schall (2004) auf allen systematisch differenzierbaren Sprachebenen (1) die Sprache der Gesetze, (2) die Sprache vor Gericht und schließlich (3) die Sprache der Täterinnen und Täter und damit verbunden die Autoren- und Sprecheridentifikation. Für Linguistinnen und Linguisten ergeben sich im skizzierten Gegenstandsbereich die Arbeitsfelder der Erforschung der Wort- und Äußerungsbedeutung, der Ähnlichkeit im Hinblick auf markenrechtliche Themen (lautlich, bildlich-schriftlich bzw. sinngemäß ähnliche Gestaltung), Beschreibung und Differenzierung einzelner Sprechakte (Beleidigung, Lügen usw.), Auseinandersetzung mit Produktbeschreibungen (Warnhinweise usw.) sowie der Autorschaftsnachweis und die

Final edited form was published in "Der Deutschunterricht". 2019, (1), S. 87-96. Friedrich Verlag. ISSN 0340-2258.

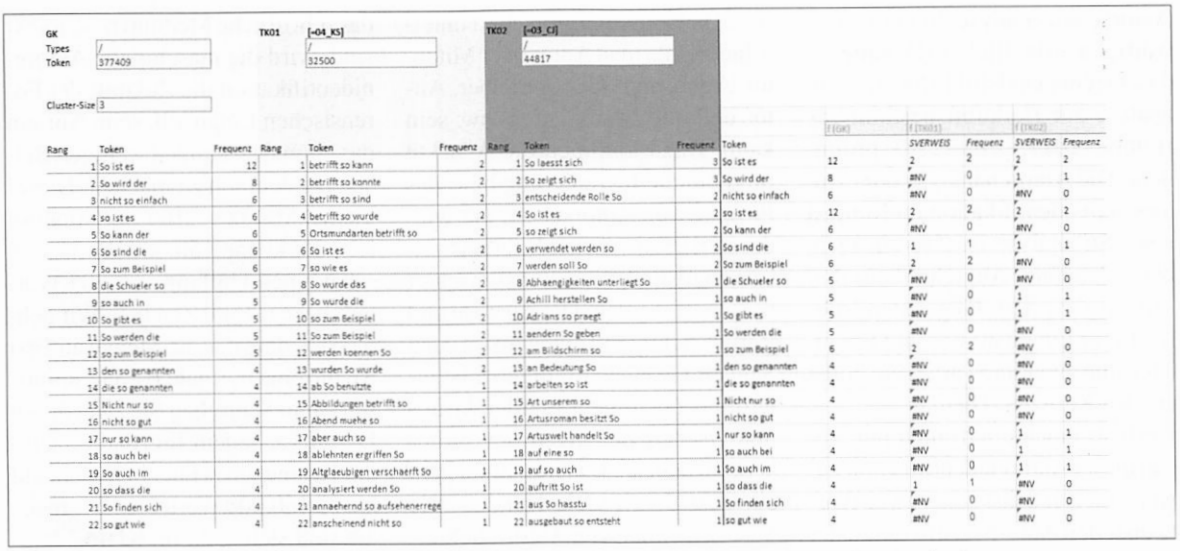
Autorschaftsanalyse. In der gegenwärtigen öffentlichen Debatte ist das Gegenstandsfeld der Sprachanalyse zur Herkunftsbestimmung (Fobbe 2017, 276–278) prominent. Die Autorenanalyse oder Autorschaftsidentifikation im Kontext einer Stilanalyse ist also ein recht überschaubarer Ausschnitt aus diesem skizzierten Gegenstandsbereich; für Linguistinnen und Linguisten, die quantitativ arbeiten, bildet sie den Kern des Themas.

Deshalb möchte ich mich nun genau dieser Autorenidentifikation zuwenden. Zunächst seien zwei Worte über den Autorbegriff gesagt, der in der forensischen Linguistik nicht unerheblich von anderen Konzepten in der Literaturwissenschaft und Sprachwissenschaft abweicht. Während wir dort nicht nach *dem* Autor fragen und die Frage „Was wollte uns der Autor damit sagen?" in Ausbildungszusammenhängen sehr früh negativ konnotiert wird, stellt die forensische Linguistik nicht nur genau diese Frage, sondern sie versucht auch einen Autor zu ermitteln, der sich in der Regel erfolgreich verbergen will (vgl. dazu Fobbe 2017, 278–283). Grundsätzlich gilt, dass der Autor als Emittent des Textes aufgefasst wird, er ist das Individuum, das Urheber des Textes ist. Kurz: „Der Autor ist die Person, die den Text verfasst hat." (Fobbe 2011, 41) Allerdings ist diese Aussage bereits metaphorisch zu verstehen. Denn differenzierter heißt das: „Der Emittent des Textes ist Autor des Textes, da er die Gestaltungsmacht [...] über den Text hat und auswählt, was, wie, an welcher Stelle und in welcher Form Teil des Textes wird." (Fobbe 2011, 42) Ganz konkret werden in der Forensischen Linguistik singuläre und multiple (bzw. kollektive) Autorschaft unterschieden, da man Autor oder Mitautor, Autor und/oder Schreiber, Autor und/oder Kompilator usw. sein kann. Dazu kommen die Aspekte rund um die Inszenierung bzw. das Postulat von Autorschaft, die wieder stärker auch in der Literaturwissenschaft und Sprachwissenschaft diskutiert werden. Die Analyse und Berücksichtigung eines auktorialen Selbstbilds eines Emittenten ist allerdings noch nicht in allen kriminologischen Zugängen tief verankert. So oder so: Im Normalfall geht ein Text durch viele Hände – die Vervielfachung von Autor-, Schreiber- und Herausgeberschaft hat eine „Schutzfunktion", sie soll den Autoren verbergen.

Bei der Analyse der Autorschaft stehen typische Autorenmerkmale auf allen Sprachsystemebenen zur Debatte und je nach Quellenart des emittierten Textes sind drei Felder der Untersuchung zu unterscheiden. Zum einen ist die Sprecheridentifikation zu nennen (Phonetik). Einen zweiten großen Bereich bildet die Handschriftenanalyse und Graphologie, die allerdings äußerst fehleranfällig ist, und meist nur mit Spuren einer Quelle arbeiten kann. Der bekannteste Fall in der jüngeren Geschichte ist der so genannte „Kreuzworträtselmord", in dem eine Reihe graphologischer Merkmale in begonnenen Kreuzworträtseln schlussendlich zur Überführung des Täters führten. Dieser Fall ist in seiner Form absolut einzigartig und auch heute noch überaus faszinierend. Das dritte Feld ist die maschinelle Autorenanalyse. Da heutzutage alle kaum übersehbare digitale Spuren hinterlassen, die mittlerweile dank Sprachidentifikationsdiensten nicht mehr nur auf das schriftliche Medium beschränkt sind, wird die maschinelle Autorenidentifikation die Zukunft der Forensischen Linguistik sein. Nur um das einmal an einigen wenigen Beispielen deutlich zu machen: Dieser kleine Absatz – „Bei der Analyse [...] auf knapp 150.000 Wörter an" – hat einen Umfang von 205 Wörtern bei ca. 1500 Zeichen. Wir nehmen an, dass das die Menge an Text ist, die eine normale E-Mail umfassen kann. Schreiben Sie zwei dieser E-Mails an jedem Tag, dann wächst die Textmenge in nur einem Jahr auf knapp 150.000 Wörter an. Die meisten sind sich nicht im Klaren darüber, wie präzise sich bereits Vorannahmen zur Autorschaft zu einem Autorenstil aus dieser Textmenge ableiten lassen – entsprechende Vergleichsdaten vorausgesetzt. Nach zehn Jahren haben Sie ein Autorschaftskorpus mit 1.500.000 Wörtern aufgebaut, das statistischen Signifikanztests standhält – mit je nur zwei E-Mails am Tag; Online-Bestellungen, Aktivitäten im Web 2.0 (Blog, Facebook, Twitter, Pinterest, Instagram), Chat-messages (WhatsApp, SnapChat, Telegram) und Sprachäußerungen in Gegenwart eines iOS- oder Android-Smartphones, Sprachidentifikationstechnik im Smart Home (Alexa, Google Home usw.) noch nicht mitgerechnet.

Die maschinelle Analyse zur Ermittlung eines Autorenstils und damit einer Zuordnung inkriminierter Texte zu einem Autoren mittels statistischer Verfahren setzt genau an diesem Punkt an. Sie kann große Mengen an Texten (TK01 und TK02 in Abb. 1) nach spezifischen Merkmalen (hier: Dreiwortgruppen / Trigrammen) zu einer anderen Textmenge (GK in Abb. 2) in Relation

Final edited form was published in "Der Deutschunterricht". 2019, (1), S. 87-96. Friedrich Verlag. ISSN 0340-2258.



GK		TK01	[-04_KS]	TK02	[-03_CJ]							
Types	/		/		/							
Token	377409		32500		44817							
Cluster-Size	3											
Rang	Token	Frequenz	Rang	Token	Frequenz	Rang	Token	Frequenz	Token	f (GK)	f (TK01) SVERWEIS	Frequenz	f (TK02) SVERWEIS	Frequenz
1	So ist es	12	1	betrifft so kann	2	1	So laesst sich	3	So ist es	12	2	2	2	2
2	So wird der	8	2	betrifft so konnte	2	2	So zeigt sich	3	So wird der	8	#NV	0	1	1
3	nicht so einfach	6	3	betrifft so sind	2	3	entscheidende Rolle So	2	nicht so einfach	6	#NV	0	#NV	0
4	so ist es	6	4	betrifft so wurde	2	4	So ist es	2	so ist es	12	2	2	2	2
5	So kann der	6	5	Ortsmundarten betrifft so	2	5	so zeigt sich	2	So kann der	6	#NV	0	#NV	0
6	So sind die	6	6	So ist es	2	6	verwendet werden so	2	So sind die	6	1	1	1	1
7	So zum Beispiel	6	7	so wie es	2	7	werden soll So	2	So zum Beispiel	6	2	2	#NV	0
8	die Schueler so	5	8	So wurde das	2	8	Abhaengigkeiten unterliegt So	1	die Schueler so	5	#NV	0	#NV	0
9	so auch in	5	9	So wurde die	2	9	Achill herstellen So	1	so auch in	5	#NV	0	1	1
10	So gibt es	5	10	so zum Beispiel	2	10	Adrians so praegt	1	So gibt es	5	#NV	0	1	1
11	So werden die	5	11	So zum Beispiel	2	11	aendern So geben	1	So werden die	5	#NV	0	#NV	0
12	so zum Beispiel	5	12	werden koennen So	2	12	am Bildschirm so	1	so zum Beispiel	6	2	2	#NV	0
13	den so genannten	4	13	wurden So wurde	2	13	an Bedeutung So	1	den so genannten	4	#NV	0	#NV	0
14	die so genannten	4	14	ab So benutzte	1	14	arbeiten so ist	1	die so genannten	4	#NV	0	#NV	0
15	Nicht nur so	4	15	Abbildungen betrifft so	1	15	Art unserem so	1	Nicht nur so	4	#NV	0	#NV	0
16	nicht so gut	4	16	Abend muehe so	1	16	Artusroman besitzt So	1	nicht so gut	4	#NV	0	#NV	0
17	nur so kann	4	17	aber auch so	1	17	Artuswelt handelt So	1	nur so kann	4	#NV	0	1	1
18	so auch bei	4	18	ablehnten ergriffen So	1	18	auf eine so	1	so auch bei	4	#NV	0	1	1
19	So auch im	4	19	Altglaeubigen verschaerft So	1	19	auf so auch	1	So auch im	4	#NV	0	1	1
20	so dass die	4	20	analysiert werden So	1	20	auftritt So ist	1	so dass die	4	1	1	#NV	0
21	So finden sich	4	21	annaehernd so aufsehenerrege	1	21	aus So hasstu	1	So finden sich	4	#NV	0	#NV	0
22	so gut wie	4	22	anscheinend nicht so	1	22	ausgebaut so entsteht	1	so gut wie	4	#NV	0	#NV	0

Abb. 1: Vergleich von zwei Teilkorpora (TK01 und TK02) in Relation zu einem Gesamtkorpus (GK) im Hinblick auf lexikalisch spezifizierte Trigramme (umgesetzt mit MS Excel)

setzen und vergleichen. Um bei den schon erwähnten Zahlen zu bleiben: Faktisch reichen 100.000 Token von elf Personen aus, um statistisch signifikante Merkmale eines Autorenstils zu erarbeiten. Es können in der maschinellen Analyse Formmerkmale (Vokal- und Konsonantengraphem-Relationen, Wort- und Satzlängen), Worthäufigkeiten, Kollokationen (signifikantes, gemeinsames Auftreten von Wörtern und grammatischen Wortarten), Kollostruktionen (signifikantes gemeinsames Auftreten von Konstruktionen) und Fehler in Orthografie, Grammatik und Aussprache (Fremdwortphonologie) verglichen werden. Die Vorteile liegen auf der Hand: Sprachlich auch den Emittenten unbewusste Muster können durch maschinelle Analysen herausgearbeitet werden, die Autorschaft kann auf der Basis unterschiedlichster Merkmale mit sehr hoher Wahrscheinlichkeit anhand eines signifikanten Stils ermittelt werden.

4 Möglichkeiten der maschinellen Stilanalyse und Autorenidentifikation und Stilanalyse im Deutschunterricht: Was ist das besondere an Goethes *Faust*?

Man kann das Set an Tools und an Werkzeugen und an Fragen auch verwenden, ohne, dass es auf den ersten Blick um die Ermittlung von Autorschaft geht, sondern dafür, einen bestimmten Autorenstil zu erarbeiten (vgl. im knappen Überblick Fobbe 2017, 280–283 und vor allem die in Kap. 2.1 vorgestellte angloamerikanische Forschung). Leitend soll die Frage sein, was korpuslinguistische und damit auch statistische Methoden[3] leisten können, wenn man sie nutzt, um Hypothesen für eine qualitative Untersuchung und Analysen zu generieren und z. B. Interpretationen literarischer Texte auf eine empirisch fundierte Datenbasis zu stellen – das man dabei ein ganzes Set an Tools nutzen kann, mit dem man in anderen Kontexten Autorschaftsidentifikationen durchführt, liegt auf der Hand. Als forensischer Linguist kann man offen fragen, welche Konsequenzen es hätte, wenn man sprachliche Besonderheiten literarischer Texte einfach errechnete, um Interpretationsansätze zu entwickeln oder vorliegende Interpretationen zu bestätigen oder zu korrigieren. Hier will ich exemplarisch folgende Fragen stellen: Welche sprachlichen Muster in Goethes *Faust*-Dramen unterscheiden diese *Texte* (signifikant) von anderen Dramen? Statt einer voraussetzungslosen Lektüre setzen wir auf eine hypothesengestützte maschinelle Auswertung vorliegender (exemplarisch ausgewählter) Dramentexte und verwenden dafür frei verfügbare Fassungen (z. B. von www.digbib.org/, 8.6.2018):

Gotthold Ephraim Lessing (1729–1781): *Minna von Barnhelm* (1767), *Emilia Galotti* (1772), *Nathan der Weise* (1779);

Final edited form was published in "Der Deutschunterricht". 2019, (1), S. 87-96. Friedrich Verlag. ISSN 0340-2258.

Friedrich von Schiller (1759–1805): *Die Räuber* (1781), *Wilhelm Tell* (1804);
Heinrich von Kleist (1777–1811): *Penthesilea* (1808), *Der zerbrochene Krug* (1811), *Prinz Friedrich von Homburg* (1821);
Johann Wolfgang von Goethe (1749–1832): *Faust. Eine Tragödie* (1808), *Faust. Der Tragödie zweiter Teil* (1832);
Franz Grillparzer (1791–1872): *Der Traum ein Leben* (1840);
Heinrich Heine (1797–1856): *Almansor* (1821)

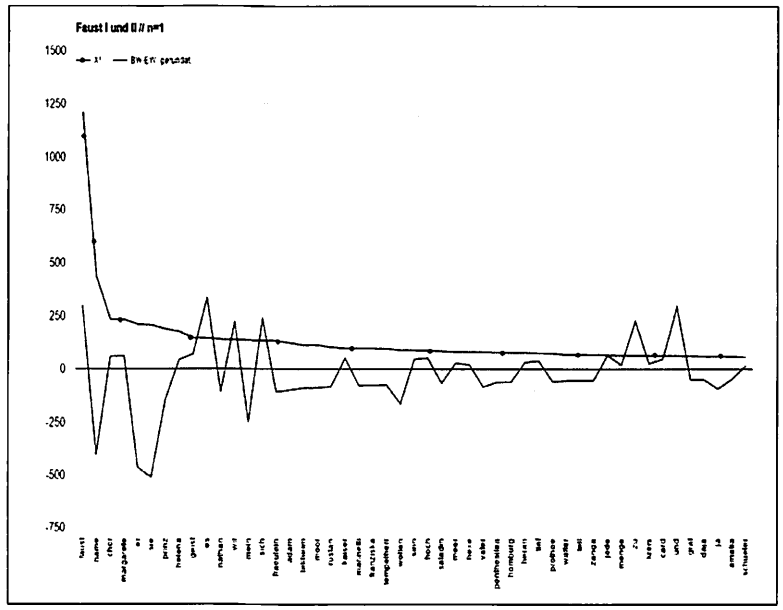


Abb. 2: Goethes *Faust*-Texte in Relation zu anderen Dramentexten, lexikalische Einheiten (umgesetzt mit MS Excel)

Zur Berechnung setzen wir auf Tools, die (fast) alle frei verfügbar sind.[4] Das sind der

- *AntFileConverter*: Dieses Tool hilft dabei, Texte in den gängigen Formaten (*.pdf, *.doc[x] und *.txt) für die maschinelle Analyse vorzubereiten,
- *TagAnt*: Dieses Tool wird verwendet, um Korpora zu lemmatisieren (also Flexionsdifferenzen zu normalisieren) und grammatisch zu annotieren,
- *AntConc*: Mit AntConc können sprachliche Muster unterschiedlicher Strukturierung gezählt und verrechnet werden. Wir werden es hier dazu nutzen, um n-Gramme auf lexikalischer und auf grammatischer Basis zählen zu lassen, und
- *Excel* (oder eine alternative Tabellenkalkulation): Berechnung von Signifikanzprofilen.

Mit diesen Hilfsmitteln, die im Unterricht auf basaler Rechnertechnik eingesetzt und genutzt werden können, und nach der Vorbereitung des Korpus errechne ich so genannte Signifikanzprofile, die Hinweise auf einen möglichen Autorenstil oder die Spezifik eines bestimmten Textes geben. Signifikanztests können mit unterschiedlichen statistischen Maßen errechnet werden – für diesen ersten Zugang wird der so genannte *Chi Quadrat*-Test eingesetzt. Kurz gesagt vergleicht man zwei (oder mehr) Korpora geleitet in Bezug auf das Vorkommen bestimmter sprachlicher Muster ausgehend von der Hypothese, dass die Verteilung sprachlicher Muster zufällig ist. Das können Einzelwörter sein (z. B. *Faust*), Mehrworteinheiten *(des Pudels Kern)* oder grammatische Konstruktionen (NOMEN mit VERB mit ADJEKTIV). Bestätigt sich die Nullhypothese nicht, dann ist die Verteilung (und damit die Korrelation sprachlicher Muster) mit einer bestimmten Wahrscheinlichkeit nicht zufällig – ein untersuchtes Korpus weicht also bzgl. bestimmter Merkmale von der erwarteten Verteilung ab und lässt einen Autoren- bzw. Textstil sichtbar und mehr oder weniger wahrscheinlich werden. Im konkreten Beispiel: Der statistische Signifikanztest sagt aus, welche sprachlichen Muster auf einer bestimmten sprachlichen Ebene als besonders typisch für einen Text oder Autorenstil zu gelten haben, zugleich aber auch, ob deren *Auftreten* (z. B. *Faust*) oder *Fehlen* (z. B. *Prinz*) typisch für die Goethe'schen *Faust*-Texte im Vergleich zu den anderen Texten ist. Diese zusätzlichen Aussagen können aus der Differenz zwischen beobachteten und erwarteten Werten abgeleitet werden, die man dafür aus der Signifikanzberechnung zweitverwerten kann (vgl. dazu oben Abb. 1).

Beginnen wir mit den 50 auffälligsten, lemmatisierten, lexikalischen Einheiten (Abb. 2) – das sind die, die man auch durch sorgfältige Lektüre ermitteln könnte. Alle

Final edited form was published in "Der Deutschunterricht". 2019, (1), S. 87-96. Friedrich Verlag. ISSN 0340-2258.

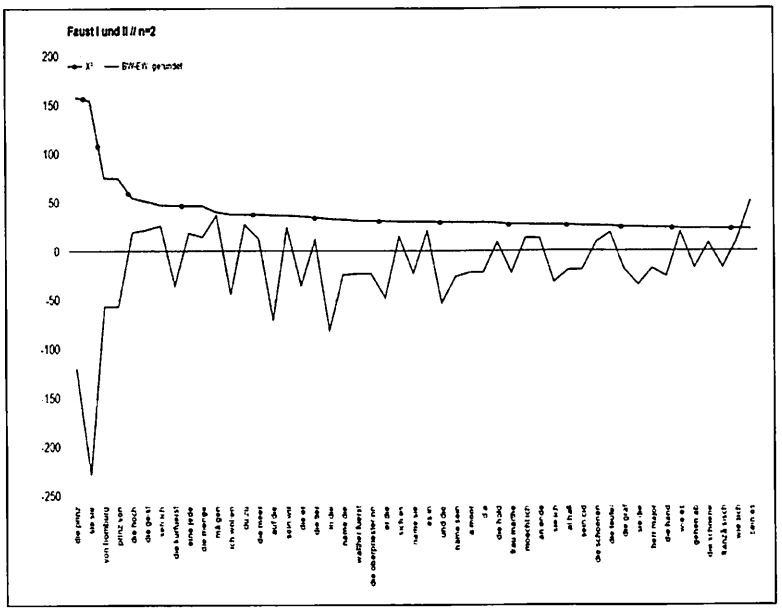


Abb. 3: Goethes *Faust*-Texte in Relation zu anderen Dramentexten, lexikalische Bigramme (umgesetzt mit MS Excel)

angezeigten Werte in der abfallenden Kurve des Chi-Quadratwertes sind statistisch höchstsignifikant. Sie sind also mit einer nahezu huntertprozentigen Wahrscheinlichkeit *nicht zufällig*, d. h. *typisch* für die *Faust*-Dramen. Die zweite Kurve weist die Differenz zwischen beobachteten und erwarteten Werten aus. Ist der Wert für eine Einheit positiv, dann ist ihr *Auftreten* besonders typisch für die *Faust*-Texte; bei negativem Wert ist ihr *Fehlen* typisch. Wenig überraschend ist, dass *Faust* als lexikalische Einheit überproportional häufig und statistisch signifikant verwendet wird. Neben Trivialitäten wie dieser sind hier an ‚Einzelwortschicksalen' das häufige Auftreten von *es*, *sich* als Reflexivmarker oder der Konjunktion *und* sowie das Ausbleiben erwarteter Personalpronomina der dritten Person (*er* und *sie*) auffällig. Letzteres – dies eine erste Hypothese – könnte möglicherweise Indikator für die besondere Dialogizität der Goethe'schen Dramen sein, wofür stützend auch das hohe Auftreten zumindest des Personalpronomens *wir* spricht.

Das gilt auch für eine ganze Reihe lexikalischer, lemmatisierter Zweiworteinheiten (Bigramme, n=2, in Abb. 3), die ebenso automatisch errechnet werden. Von durch Handarbeit zu vermeidenden Fehlern (wie *mögen* auf Position 11) oder nicht vermeidbaren Fehlern (wie der Doppelangabe von *sie*l*Sie* auf Position 2) abgesehen, stellt sich schon bei den Bigrammen der für lexikalische Einheiten typische Effekt ein, dass Eigennamen *(der Prinz, von Homburg, Prinz von)* eine dominante Rolle spielen. Die Beispiele referieren auf Kleist und freilich lassen sie sich in Goethes Texten nicht nachweisen. Interessant sind Bigramme wie *eine jede* (auf Position 9) *sein es* (auf der letzten Position), die man nicht zwingend mit Goethes Texten in Verbindung bringen würde und die vermutlich einer genauen Lektüre auch entgingen – Beobachtungen wie diese wären in einem zweiten Schritt am Originaltext zu prüfen, um mögliche Hypothesen für eine Interpretation aufzubauen. Erweitert man den Zugriff auf lexikalische, lemmatisierte Trigramme (n=3), kommt man an den Rand der Aussagekraft dieser Auswertungsmöglichkeit. Dennoch lassen sich Vorlieben für bestimmte sprachliche Muster noch entdecken, deren genaue Interpretation man im Unterricht weiterverfolgen könnte: *Da liegen die* für z. B. *da liegt der*, *da liegen die*, *da lag das* usw. ist als lemmatisiertes Muster eine für Goethes Texte signifikant und häufiger als erwartet auftretende Mehrworteinheit.

Spannend wird es noch einmal, wenn man die Korpusdaten grammatisch annotiert – also z. B. Wortartenannotationen vornimmt und diese in unterschiedlich definierten Umgebungen untersucht (n=1 bis n=3). Ich kürze den Weg hier ab und stelle in diesem Beitrag nur die Trigramme aus dem grammatisch annotierten Korpus vor, die sich als typisch für Goethes Dramen erweisen.

Diese Herangehensweise hat den Vorteil, dass man in der maschinellen Analyse von der lexikalischen Ebene abstrahiert. So fallen Muster (und deren Kombinationen in Signifikanzprofilen) auf, die in der forensischen Analyse treffsicher Aussagen über Autorenstile erlauben. Auf das Dramenbeispiel angewendet müsste man diese Schlüsse freilich abwandeln: Ich habe, ge-

Final edited form was published in "Der Deutschunterricht". 2019, (1), S. 87-96. Friedrich Verlag. ISSN 0340-2258.

wissermaßen in einer Potentialanalyse als Anregung für den Deutschunterricht, Goethes Dramen-Stil (exemplarisch am *Faust*) im Vergleich zu anderen Dramen ermittelt, ohne zugleich aber Aussagen über *einen* Autorenstil generell machen zu können oder für *alle* Texte Goethes zu sprechen usw. Aus den Ergebnissen will ich zwei Aspekte in den Mittelpunkt rücken: So genannte Zwillingsformeln (NN KON NN / NOMEN mit KONJUNKTION mit NOMEN) auf Position 2 in Abb. 4 und der häufige Gebrauch von attributiven Adjektiven (ADJA/D NN / ADJEKTIV mit NOMEN) auf den Positionen 3, 4, 8, 13 und 14 zeichnen die Dramentexte Goethes im Vergleich zu Dramentexten seiner Zeitgenossen aus; typischerweise werden dementsprechend Nominalphrasen mit Artikel und Nomen ohne attributives Adjektiv äußerst selten verwendet (vgl. Positionen 12 und 30 in Abb. 4). Auch das gehört vor dem Vergleichshintergrund der anderen Dramen zur Typik dieser Texte – spätestens das letzte Ergebnis, nämlich auch zu sagen, *welche Einheiten typischerweise nicht in einem Text zu erwarten sind*, kann man auf dieser Ebene, der Untersuchung grammatisch annotierter Korpora, nicht mehr durch Lektüreerfahrung erzielen.

Blicken wir abschließend noch auf einzelne Belege, die wir anhand dieser exemplarischen grammatischen Muster als besonders typisch und häufig in Goethes *Faust*-Dramen ausgemacht haben. Schon in der Zueignung in Goethes *Faust* I findet man bspw. Zwillingsformeln wie diese aus dem horizontal annotierten Korpus

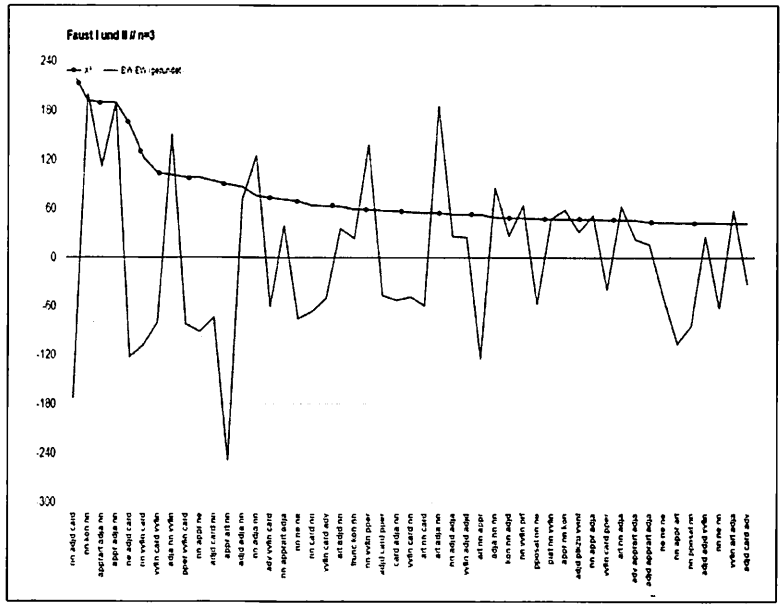


Abb. 4: Goethes *Faust*-Texte in Relation zu anderen Dramentexten, Trigramme grammatisch annotierter Einheiten (umgesetzt mit MS Excel)

*Dunst*_NN *und*_KON *Nebel*_NN
*Lieb*_NN *und*_KON *Freundschaft*_NN
*Not*_NN *und*_KON *Truebsal*_NN

Und adjektivische Attribuierungen durchziehen beinahe so ikonisch wie die Zwillingsformeln den gesamten Text (stellvertretend hier auch aus der Zueignung):

*Ihr*_PPER *naht*_VVFIN *euch*_PPER *wieder*_ADV ,_$, ***schwankende***_ADJA ***Gestalten***_NN ,_$, *die*_ART *frueh*_ADV *sich*_PRF *einst*_ADV *dem*_ART ***trueben***_ADJA ***Blick***_NN *gezeigt*_VVPP ._$.

Natürlich ist mit deutlichen Ergebnissen wie diesen auch der Umkehrschluss zulässig, der nicht auf Mutmaßungen aufruht: Sie stoßen in einem Drama von den hier verglichenen Autoren nicht auf Zwillingsformeln und es werden kaum attributive Adjektive verwendet? Dann ist das Drama nicht von Goethe (bzw. genauer: nicht die *Faust*-Dramen, alles andere wäre noch zu prüfen).

5 Fazit und Ausblick

Der forschungspraktische Vorteil der maschinellen Stilanalyse und Autorschaftsidentifikation liegt sofort auf der Hand: Ohne Vorannahmen und Kenntnis von kultur- und sprachhistorischen Voraussetzungen, zum Dramenaufbau oder der allgemeinen Charakteristik dramatischer Texte (wie Dialogizität statt Narrativität) kann man sprachliche Besonderheiten ausgesuchter Texte zuverlässig ermitteln. Und zwar so genau, wie es Lektüre niemals könnte. Aber: Die Ergebnisse sind offener Natur und immer interpretationsbedürftig. Die Daten spre-

chen nicht für sich und deshalb beginnt hier die interpretative Arbeit, auch wenn ein typisches Stereotyp über Linguistinnen und Linguisten behauptet, dass diese nur Daten auszählen und sich damit begnügen. Der Zugang ist besonders geeignet, um im Deutschunterricht linguistische und literaturwissenschaftliche Perspektiven zu verzahnen und Schülerinnen und Schüler, die sich eher für naturwissenschaftliche Fragestellungen interessieren, mit einer Methode vertraut zu machen, die sie anwenden können, um Texte auf eine spezielle Weise aufzuschließen. Die gewonnenen Daten können Grundlage eigener Hypothesen und Interpretationszugänge werden. Zum Beispiel ist nach der besonderen Rolle der Zwillingsformeln in Goethes Texten ebenso zu fragen wie nach der Bedeutung von attributiv gebrauchten Adjektiven in seinen Dramentexten oder besonderen lexikalischen Bi- und Trigrammen (*und jede*, *da liegen die* usw.). Und, diese Frage habe ich nur angedeutet: Stellen Goethes Texte eine besondere Dialogizität aus, die anderen Dramentexten der Zeit nicht eingeschrieben ist? Löst man sich vom Beispiel dieses Artikels, dann lassen sich Fragen formulieren wie diese: Kann man ein Autorschaftskorpus anhand statistischer Analysen in verschiedene Phasen einteilen? Kann man bisher anonyme Texte mittels maschineller Analysen einem Autorenkorpus zuweisen? Wie und wenn ja auf welche Weise unterscheiden sich Romane des Realismus und Naturalismus? Weisen Gedichte eines Autors eine spezifische Textur auf, die nur auf der Basis grammatischer Annotation sichtbar wird? Welche Beschreibungsmöglichkeiten eröffnen sich, wenn alternative Korpora (etwa vom *Digitalen Wörterbuch der deutschen Sprache*, http://dwds.de, oder vom *Deutschen Textarchiv*, http://deutsches-textarchiv.de, Stand: 8.6.2018) im Vergleich zwischen literarischen und wissenschaftlichen Texten z. B. des 19. Jahrhunderts hinzugezogen werden? Wie singulär sind serielle Schilderungen (vgl. auch Bubenhofer 2018) oder was sind typische Konstruktionen narrativer Texte (vgl. Ziem/Lasch 2018 sowie Lasch 2018)? ▩

Literatur

Online-Quellen und Tools

Analysierte Dramentexte: www.digbib.org/ [8.6.2018].

Anthony, Laurence // Ant-Tools: http://laurenceanthony.net/software.html [8.6.2018].

Bubenhofer, Noah: Einführung in die Korpuslinguistik, https://bubenhofer.com/korpuslinguistik/kurs/ [8.6.2018].

Deutschen Textarchiv: http://deutsches-textarchiv.de [8.6.2018].

Digitales Wörterbuch der deutschen Sprache: http://dwds.de [8.6.2018].

Hewitt, Mark Algee, Stanford Literary Lab: https://litlab.stanford.edu/ [8.6.2018].

Lasch, Alexander, Sprachpunkt: Forensische Linguistik: https://alexander-lasch.wordpress.com/?s=Forensische+Linguistik&x=0&y=0 [8.6.2018].

Mephisto und Faust beim Schach (19. Jh.), anonym https://goo.gl/Q3pam6 [8.6.2018].

Scharloth, Joachim, Surveillance and Security: Authorship Identification: http://security-informatics.de/blog/ [8.6.2018].

Sekundärliteratur

Balossi, Giuseppina (2014): A Corpus Linguistic Approach to Literary Language and Characterization. Virginia Woolf's *The Waves* (Linguistic Approaches to Literature 18).

Biber, Douglas (2011): Corpus linguistics and the study of literature? In: Scientific Study of Literatur 1(1), 15–23.

Biber, Douglas/Randi Reppen (Hg.) (2015): The Cambridge Handbook of English Corpus Linguistics (CHECL). Cambridge.

Bubenhofer, Noah (2018): Serialität der Singularität. In: LiLi 48(2), 357–388.

Dern, Christa (2009): Autorenerkennung. Theorie und Praxis der linguistischen Tatschreibenanalyse. Stuttgart.

Fischer-Starcke (2010): Corpus Linguistics and Literature. Corpus Stylistic Analyses of Literary. Works by Jane Austen and her Contemporaries. London.

Fobbe, Eilika (2011): Forensische Linguistik. Eine Einführung. Tübingen.

Fobbe, Eilika (2017): Forensische Linguistik. In: Ekkehard Felder/Friedemann Vogel (Hg.): Handbuch Sprache im Recht (HSW 12). Berlin, Boston, 271–289.

Genette, Gérard (2010): Die Erzählung (frz. 1972/1983). 3. Aufl. München.

Herrmann, J. Berenike & Gerhard Lauer (2018): Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne. In: Osnabrücker Beiträge zur Sprachtheorie 92.

Hoover, David L./Jonathan Culpeper/Kieran O'Halloran (2014): Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama (Routledge Advances in Corpus Linguistics 16). New York, London.

Keshabyan, Irina/Ángela Almela (Hg.) (2012): A New Approach to Literature: Corpus Linguistics (Special Issue of the International Journal of English Studies 12,2).

Kniffka, Hannes (1981): Der Linguist als Gutachter bei Gericht. In: Günter Peuser/Stefan Winter (Hg.): Angewandte Srachwissenschaft: Grundfragen, Berichte, Methoden. FS Günter Kandler. Bonn, 584–634.

Kniffka, Hannes (1994): Ein Gutachten über ein Fehlgutachten. In: Kriminalisitik und forensische Wissenschaften 82, 111–130.

Kniffka, Hannes (2001): Eine Zwischenbilanz aus der Werkstatt eines ‚forensischen' Linguisten: Zur Analyse autonymer Autorschaft. In: Linguistische Berichte 185, 75–104.

Kniffka, Hannes (2007): Working in Language and Law. A German perspektive. Basingstoke, New York.

Kniffka, Hannes (Hg.) (1990): Texte zu Theorie und Praxis forensischer Linguistik (Linguistische Arbeiten 249). Tübingen.

Lasch, Alexander (2018): Phrasale Konstruktionen als Basis narrativer Routinen. In: ZGL 46(1), 44–64.

Levi, Judith (1982): Linguistics, Language, and the Law. A Topical Bibliography. Bloomington,Ind.

Mahlberg (2007a): Corpus stylistics: bridging the gap between linguistic and literary studies. In: Michael Hoey et al. (Hg.): Text, Discourse and Corpora: Theory and Analysis. London, 219–246.

Mahlberg, Michaela (2007b): Clusters, key clusters and local textual functions in Dickens. In: Corpora 2(1), 1–31.

Mahlberg, Michaela (2013): Corpus Stylistics and Dickens's Fiction (Routledge Advances in Corpus Linguistics 14). New York/London.

Mahlberg, Michaela (2015): „Literary style and literary texts". In: Douglas Biber/Randi Reppen (Hg.) (2015): The Cambridge Handbook of English Corpus Linguistics (CHECL). Cambridge, 346–361.

Moretti, Franco (2017): Patterns and Interpretation (Literary Lab Pamphlets 15). Online verfügbar unter: https://litlab.stanford.edu/LiteraryLabPamphlet15.pdf [8.6.2018].

Schall, Sabine (2004): Forensische Linguistik. In: Karlfried Knapp (Hg.): Angewandte Linguistik. Ein Lehrbuch. Tübingen, 455–562.

Spieß, Constanze/Doris Tophinke (Hg.) (2018): Alltagspraktiken des Erzählens (Sonderheft LiLi 48,2).

Ziem, Alexander/Alexander Lasch (2018): Konstruktionsgrammatische Zugänge zu narrativen Texten. In: LiLi 48(2), 389–410.

Anmerkungen

(1) Vgl. dazu den 2018 erschienen thematischen Band der *Zeitschrift für Literaturwissenschaft und Linguistik* (LiLi) zu „Alltagspraktiken des Erzählens" mit Beiträgen von Spieß/Tophinke, Quasthoff/Stude, Gredel/Mell, König/Oloff, Kotthoff, Hoffmann, Weidacher. Für den hier skizzierten Zusammenhang sind besonders die Beiträge von Bubenhofer 2018 sowie Ziem/Lasch 2018 relevant.

(2) Das skizzierte Verhältnis lässt sich exemplarisch am Blog „Surveillance and Security" von Joachim Scharloth nachvollziehen, der sich der Autorschaftsidentifikation zuwendet (http://security-informatics.de/blog/?cat=22, 8.6.2018), das Thema aber nicht in Richtung einer fundierten Forensischen Linguistik treibt, sondern es aus der Perspektive eines Korpuslinguisten als Problemstellung aufnimmt. Gleiches gilt für mein Blog „Sprachpunkt" (https://alexanderlasch.wordpress.com/?s=Forensische+Linguistik, 8.6.2018): In der akademischen Lehre und in kleinen Modellanalysen illustriere ich, was man mit korpuslinguistischen Tools zu zeigen im Stande ist; eine institutionalisierte Forensische Linguistik wird damit aber ebenfalls nicht vorangetrieben.

(3) Für einen Einstieg in das Thema sei Noah Bubenhofers Online-*Einführungskurs in die Korpuslinguistik* empfohlen (https://bubenhofer.com/korpuslinguistik/kurs/, 8.6.2018) empfohlen.

(4) Die drei Korpustools stellt Laurence Anthony frei verfügbar für Windows, macOS und Linux zur Verfügung (http://laurenceanthony.net/software.html, 8.6.2018)

Final edited form was published in "Der Deutschunterricht". 2019, (1), S. 87-96. Friedrich Verlag. ISSN 0340-2258.

Autorinnen und Autoren

Prof. em. Dr. Gerd Antos war von 1993 bis 2014 Professor für „Germanistische Linguistik" an der Martin-Luther-Universität in Halle-Wittenberg. Seine Arbeitsgebiete sind die Angewandte Linguistik, Formulierungs- bzw. Schreibforschung, Laienlinguistik, sprachlicher Wissenstransfer (Experten-Laien-Kommunikation), Rechtslinguistik, Verständlichkeitsforschung, Sprache in der Digitalkultur.

PD Dr. habil. Franz d'Avis lehrt als Akademischer Oberrat am Deutschen Institut der Johannes Gutenberg-Universität Mainz. Seine Forschungsinteressen liegen im Bereich der Syntax, Semantik und Pragmatik und ihren Schnittstellen. Im Rahmen der Lehrerausbildung schließt das auch Fragen zur Didaktik in diesen Bereichen mit ein.

Prof. Dr. Matthias Ballod lehrt Didaktik der Deutschen Sprache und Literatur an der Martin-Luther-Universität Halle-Wittenberg. Seine Forschungsschwerpunkte liegen im Bereich der Wissenskommunikation und Informationsdidaktik, digitale Medien in Lehr-Lernkontexten von Schule und Hochschule, interkultureller Wissenstransfer (DaF/DaZ).

Dr. Katharina Böhnert ist wissenschaftliche Mitarbeiterin am Lehr- und Forschungsbereich Fachdidaktik Deutsch der RWTH Aachen University. Ihre Forschungsschwerpunkte liegen im Bereich des sprachreflexiven Deutschunterrichts, insbesondere unter diachroner Perspektive.

Prof. Dr. Rita Finkbeiner lehrt Germanistische Sprachwissenschaft an der Heinrich-Heine-Universität Düsseldorf. Ihre Forschungsschwerpunkte liegen in der Grammatik und Pragmatik der deutschen Gegenwartssprache, insbesondere in den Bereichen Semantik/Pragmatik-Schnittstelle, Syntax/Pragmatik-Schnittstelle und Konstruktionsgrammatik.

Marcel Fladrich M. Ed. arbeitet als wissenschaftlicher Mitarbeiter an der Universität Hamburg und der Westfälischen Wilhelms-Universität Münster. Seine Forschungsschwerpunkt umfassen u. a. die Erforschung interaktionaler gesprochener und geschriebener Sprache sowie die Umsetzung dieser Forschungsergebnisse v. a. im Deutsch-als-Fremdsprache-Unterricht.

Prof. Dr. Wolfgang Imo lehrt germanistische Linguistik an der Universität Hamburg. Seine Forschungsschwerpunkte umfassen u. a. die Grammatik des Deutschen, die Erforschung interaktionaler gesprochener und geschriebener Sprache sowie die Umsetzung dieser Forschungsergebnisse v.a. im Deutsch-als-Fremdsprache-Unterricht.

Prof. Dr. Alexander Lasch lehrt germanistische Linguistik und Sprachgeschichte an der Technischen Universität Dresden (https://gls-dresden.de). Forschungsschwerpunkte sind Konstruktionsgrammatik, Digital Humanities sowie diskurs- und domänenspezifische Kommunikation in Vergangenheit und Gegenwart.

Prof. em. Dr. Bernd Müller-Jacquier lehrt und forscht im Bereich Interkulturelle Kommunikation/Deutsch als Fremdsprache; Publikationen zur interkulturellen Didaktik und Mediation mit Fokus auf Prozesse des Fremdverstehens (u. a. Wirtschaftskommunikation, Tourismus, Semiotik interkulturellen Handelns); Konzeption videogestützer Trainingsmaterialien.

Jun.-Prof. Dr. Jessica Nowak lehrt Historische Sprachwissenschaft des Deutschen an der Universität Mainz. Ihre Forschungsschwerpunkte liegen im Bereich der historisch-kontrastiven Grammatik, insbes. der diachronen Morphologie des Deutschen im Kontrast mit anderen germanischen Sprachen, in der Historischen Graphematik und Onomastik.

Sarah Stumpf M. Ed. ist wissenschaftliche Mitarbeiterin an der Martin-Luther-Universität Halle-Wittenberg. Im Projekt „[D-3] Deutsch Didaktik Digital" verantwortet sie den Arbeitsbereich Methodenentwicklung. Ihre Forschungsschwerpunkte liegen im Bereich der Sprachdidaktik, der linguistischen Pragmatik und der Spracherwerbsforschung.