colloquium: new philologies

Perspektiven einer Qualitativen Stilometrie am Beispiel stilschichtig abgesenkter Lexeme

Elisabeth Eder
Universität Klagenfurt (Austria)

Ulrike Krieg-Holz
Universität Klagenfurt (Austria)

Abstract

Here we outline an approach to stylometry, which intends to be more comprehensive compared to classical stylistic metrics and its commonly used lexical frequency counts. As a prerequisite, such an approach needs language data as a basis for its stylistic analyses. In this paper, we describe the acquisition of two relevant resources: First, we depict collecting and preparing CodE Alltag, a German-language email corpus, which contains formal expressions as well as informal and personal interactions, and thus possesses a high stylistic variability. Envisaging the analysis of the vulgar, rough or obscene dimensions of style, we then detail inducing VulGer, a lexical resource covering the lower end of the German language register.

Keywords: *Stylometry*, *Language resources*, *Emails*, *Vulgarity*

Colloquium: New Philologies, advance publication (August 2021)
doi: 10.23963/cnp.2021.6.2.2
Stable URL: https://colloquium.aau.at/index.php/Colloquium/article/view/157

1 Einleitung

Der linguistische Stilbegriff fokussiert ganz generell die Form der sprachlichen Ausgestaltung von Textstrukturen. Diese Spezifik sprachlicher Formulierungen resultiert aus der Möglichkeit, innerhalb von im Sprachsystem angelegten Varianten auszuwählen. Stil ist deshalb als ein Phänomen der Wahl anzusehen und ein Ergebnis von Entscheidungsprozessen, die sich einerseits an Vorgegebenem, Prototypischem und Musterhaftem orientieren, andererseits immer auch eigenständige Umsetzungen in Verbindung mit individualstilistischen Merkmalen darstellen. Derartige Wahlentscheidungen sind in sämtlichen Kommunikationsformen von größter pragmatischer Relevanz, weil sie das kommunikative Handeln entscheidend prägen und sowohl Textproduktions- als auch Textrezeptionsprozesse erheblich beeinflussen. Denn zum einen können sprachliche Handlungen desselben Typs auf verschiedene Weise durchgeführt werden, zum anderen unterscheiden sich sprachliche Handlungen verschiedenen Typs in stilistischer Hinsicht.

Das Ziel der sprachwissenschaftlichen Stilistik besteht darin, innerhalb von Texten und kommunikativen Zusammenhängen diejenigen Elemente und Strukturen aufzudecken, mit denen das Spezifische der sprachlichen Gestaltung einer kommunikativen Handlung charakterisiert werden kann. Dazu hat sich ein terminologisches Inventar herausgebildet, das geeignet ist, stilistische relevante Einzelelemente – etwa auf den linguistischen Beschreibungsebenen der Lexik oder Grammatik – sowie komplexere stilistische Strukturen (z.B. Stilzüge, Vertextungsstrategien) zu klassifizieren. Dieses stilistische Analyseinventar soll mit Verfahren der automatischen Textklassifikation aus dem Bereich der Computerlinguistik verbunden werden, um (semi-)automatische Stilanalysen im Sinne einer „Qualitativen Stilometrie“ zu ermöglichen.

Zum Standardinventar der klassischen Stilometrie zählen insbesondere lexikalische Häufigkeitsverteilungen und Buchstaben-N-Gramm-Frequenzen. Der Erfolg solcher Verfahren ist nicht zu bestreiten, sie operieren jedoch nur an der Oberfläche des Phänomens Stil als der spezifischen Form sprachlicher Formulierung. Im Folgenden geht es deshalb um die Entwicklung eines sogenannten Stilwerkzeugkastens, der zur automatischen Erkennung und Klassifikation stilistischer Merkmale genutzt werden kann und sich gegenüber bereits existierenden Methoden der klassischen Stilometrie stärker an der Qualität der stilistischen Einzelmerkmale orientiert. Im Zentrum des vorliegenden Beitrags steht dabei insbesondere die Erläuterung der notwendigen infrastrukturellen Voraussetzungen, da Tools dieser Art zunächst nach Sprachressourcen verlangen, auf deren Basis sprachstatistische Berechnungen durchgeführt werden können. Dazu werden zwei einschlägige Ressourcen vorgestellt: ein stilistisch breitbandig angelegtes Textkorpus (CodE Alltag) und ein deutschsprachiges Lexikon (VulGer). Letzteres repräsentiert (auf der Basis intersubjektiver Kriterien) eine spezifische Facette sprachlichen Stils,

die Vulgarität im Sinne eines stilschichtig[1] stark abgesenkten Sprachgebrauchs, und soll damit beispielhaft als Fixpunkt für empirisch valide Stilberechnungen fungieren.

Der Beitrag geht zunächst auf die besondere Problematik von E-Mail-Korpora ein und beschreibt den aktuellen Stand der Korpusarbeiten zu CodE Alltag (Erhebungsprinzipien, Entwicklungsstand, Annotation und Pseudonymisierung). Im Anschluss daran wird der methodische Ansatz der qualitativen Stilometrie erläutert und am Beispiel des abgesenkten Lexembereichs beschrieben, auf welche Weise dieser durch Verfahren der qualitativen Stilometrie präzisiert und klarer ausdifferenziert werden kann. Dazu wird eine Methode aus dem Bereich der Distributionellen Semantik – die sog. Word Embeddings – vorgestellt und auf deren Eignung für die linguistische Lexik- und Stilbeschreibung eingegangen. Schließlich wird der Aufbau des vulgärsprachlichen Lexikons (VulGer) als ein zentraler Entwicklungsschritt hin zum Stilwerkzeugkasten beschrieben.

2 Ein stilistisch breitbandiges Textkorpus: das E-Mail-Korpus CodE Alltag

Eine grundlegende Voraussetzung für eine breit angelegte Stilanalyse ist ein hohes Maß an stilistischer Varianz. Die großen deutschsprachigen Textkorpora, wie beispielsweise das DeReKo-Korpus (Kupietz & Lüngen 2014) oder das DWDS-Kern- bzw. Zeitungskorpus (Geyken 2007), erfassen primär hoch- bzw. standardsprachlich geprägte Textformen wie Zeitungstexte oder Literatur. Sie spiegeln damit nur einen Teil der gegenwärtigen Erscheinungsformen des Deutschen wider. So lassen sie etwa elektronisch vermittelte Kommunikationsformen und Textsorten, für die spezielle Ausdrucksweisen und Formulierungsmuster typisch sind, ebenso außer Acht wie die verschiedenen Formen von Alltagskommunikation. Dem trägt seit einigen Jahren eine Tendenz in der Korpuslinguistik Rechnung, die sich mit diversen Ausprägungen geschriebener informeller Alltagssprache beschäftigt, wie sie sich in Mikroblog- und Chat-Texten finden (Storrer 2013). Zu nennen ist in diesem Zusammenhang insbesondere DeRiK, ein Korpus zur Erfassung computervermittelter Kommunikation in Blogs und Chats (Beißwenger & Lemnitzer 2013). E-Mails berücksichtigt diese Textkollektion jedoch nicht. Diese Lücke schließen Aktivitäten, die am Institut für Germanistik der Universität Leipzig Mitte 2014 begonnen und an der Alpen-Adria-Universität Klagenfurt in Kooperation mit dem Institut für Germa-

[1] Die Unterteilung von Stilschichten bzw. Stilebenen ist vor allem in Zusammenhang mit der Kodifikation und Erläuterung des Wortschatzes in Wörterbüchern bekannt. Dabei wird üblicherweise von drei stilistischen Hauptebenen ausgegangen, nämlich „neutral", „gehoben" und „abgesenkt", von denen gerade die letzte weiter unterteilt wird (z.B. „umgangssprachlich", „derb", „vulgär").

nistische Sprachwissenschaft der Friedrich-Schiller-Universität Jena fortgesetzt wurden. Das Ziel dieser Arbeiten bestand im Aufbau eines umfassenden Corpus deutschsprachiger E-Mails (CodE Alltag; s.a. Krieg-Holz et al. 2016). Die Kommunikationsform *E-Mail* umfasst zahlreiche Textsorten, die inhaltlich von professionellen bis hin zu ganz persönlichen Interaktionen reichen und somit eine sehr große Bandbreite der performativen Varianz abbilden. Es zeigt sich nahezu das gesamte Kontinuum an Formulierungsmöglichkeiten. Auf der Ebene der Lexik erstreckt sich dies u.a. von stilschichtig gehoben über neutral und umgangssprachlich bis hin zu expressiven oder vulgären Elementen. Diese im Korpus enthaltene stilistische Varianz soll deshalb als Grundlage für die stilometrischen Untersuchungen dienen.

E-Mail-Korpora sind Korpora, die ausschließlich bzw. überwiegend aus elektronischer Post (E-Mails) bestehen. E-Mails treten selten isoliert auf, sondern sind häufig Teil sog. Threads, also asynchron geführter, thematisch um ein *Subject* (Betreff) zentrierter, konsekutiver Folgen von einzelnen Mails, die den Verlauf von inhaltlichen Austäuschen der Schreiber*innen dokumentieren (Krieg-Holz & Hahn 2016). Somit können solche Korpora dahingehend unterschieden werden, ob sie formal kohärent sind, weil sich Mails über ihre Thread-Struktur explizit aufeinander beziehen, oder inkohärent (ohne Kontextbezug innerhalb des E-Mail-Diskurses). Aufgrund des oft persönlichen Charakters gibt es vergleichsweise wenige solcher E-Mail-Korpora. Zudem sind im juristischen Sinne ausschließlich die Sender*innen – und eben nicht die Empfänger*innen – im Besitz des Urheberrechts an ihrer E-Mail (s.a. Krieg-Holz & Hahn 2016). Aus diesem Grund ist es auch für die linguistische Analyse notwendig, zum einen die Zustimmung der E-Mail-Produzent*innen für die Aufnahme der Texte in ein Korpus zu erlangen, zum anderen ist die Textsammlung in anonymisierter Form zusammenzustellen, um sie nachfolgend für die empirische Untersuchung nutzen zu können. Werden allerdings große Sammlungen von E-Mails von Dritten in einem eigenen, öffentlich zugänglichen Archiv zusammengetragen, die im Besitz keiner eindeutig benennbaren juristischen Person stehen, sind die Rahmenbedingungen weniger restriktiv. Dies ist etwa beim USENET-Newsgroup-Archiv der Fall, das auch innerhalb von CodE Alltag als Ressource für ein eigenes Segment genutzt wird (vgl. Krieg-Holz et al. 2016).

Die derzeit verfügbaren E-Mail-Korpora sind überwiegend englischsprachige. Hierzu gehören etwa das Enron-Korpus (Klimt & Yang 2004), das TREC (Cormack 2007), das W3C-Korpus des World Wide Web Consortiums[2] oder das Australian National Corpus (Lampert 2009). Letzteres enthält ein E-Mail-Segment (Email Australia), das mithilfe eines Aufrufs zur E-Mail-Spende aufgebaut worden ist. Für das Deutsche

[2] https://www.w3.org/

existierten vor CodE Alltag zwei E-Mail-Korpora: ein kleines von Declerck & Klein (1997), das 160 E-Mails umfasst sowie das deutlich größere FLAG-Korpus, das durch die Extraktion von 120.000 Sätzen aus der Internet USENET-Newsgroup generiert wurde (Becker et al. 2003). Angesichts des zugrundeliegenden Forschungsinteresses, der Fehlerannotation, ist das FLAG-Korpus auf sprachliche Verwendungsweisen ausgerichtet, die von standardsprachlicher Schriftlichkeit abweichen (z.B. an Mündlichkeit orientierten Schreibweisen). Demgegenüber hebt sich CodE Alltag durch eine größere sprachliche und stilistische Varianz ab (Krieg-Holz & Hahn 2016).

CodE Alltag besteht als Gesamtkorpus aus zwei gänzlich verschiedenen Partitionen. Die eine wurde auf der Grundlage des Spenden-Modells gewonnen. Sie wird auch als CodE Alltag$_{S+D}$ bezeichnet und bildet den Gegenstand der folgenden Erläuterungen. Die andere wird aufgrund ihrer Größenordnung mit CodE Alltag$_{XL}$ benannt, denn sie setzt sich aus knapp 1,5 Mio. E-Mails zusammen. Diese wurden aus sieben inhaltlich unterschiedlichen Kategorien des deutschsprachigen Teils des Internet Usenet-Newsgroup-Mail-Archivs extrahiert und minimal veredelt. Das heißt, es erfolgte die Beseitigung inhaltlicher Verrauschungen durch Fakes und Spams und gegebenenfalls die Auflösung von Zeichenkodierungsproblemen.

CodE Alltag$_{S+D}$ ist wesentlich kleiner als CodE Alltag$_{XL}$. Es besteht derzeit (Stand November 2020) aus knapp 1.500 E-Mails. Jedoch enthält es ergänzend zu den E-Mails demographische Angaben zu den E-Mail-Produzent*innen.[3] Diese Zusatzinformationen sollten ursprünglich primär für die stilistisch-forensische Analyse genutzt werden, sie sind aber darüber hinaus auch für verschiedenste variations- bzw. soziolinguistische Fragestellungen und Anwendungen nutzbar.

Beim Aufbau des CodE Alltag-Korpus wurde ein Ad-hoc-Sample angestrebt, denn eine statistisch repräsentative Auswahl von E-Mails würde nicht nur am Problem der verlässlichen Definition der Grundgesamtheit und am Privatheitsprinzip scheitern, sondern auch an der Zustimmung der Schreiber*innen zur Aufnahme ihrer E-Mails in die Stichprobe. In einem ersten Schritt wurden Germanistikstudierende in einer Einführungsvorlesung an der Universität Leipzig gebeten, jeweils eine ihrer E-Mails – ohne jegliche Veränderung der Original-E-Mail – in das Korpus zu spenden. Danach wurde über den allgemeinen Verteiler der Universität Leipzig an alle Institute und Verwaltungseinheiten ein Aufruf zur E-Mail-Spende gesendet. Um zu verhindern, dass die Schreiber*innen allesamt im universitären Umfeld tätig sind, haben wir zudem darum gebeten, auf die Spendenaktion auch im Bekanntenkreis (Familien, Freunde, Vereine usw.) hinzuweisen und so die Zusendung weiterer E-Mails zu initiieren. Die in Leipzig begonnenen Arbei-

[3] Für eine detaillierte Beschreibung beider Partitionen des CodE Alltag-Korpus vgl. Krieg-Holz et al. (2016).

ten werden seit 2015 am Institut für Germanistik der Alpen-Adria-Universität Klagenfurt fortgesetzt.[4] Hier wurden inzwischen mehrere Initiativen gestartet, um den Zulauf von E-Mails zu steigern. Das Spektrum reichte von der Verteilung von Flyern, über die persönliche Ansprache von Passant*innen in öffentlichen Räumen und Besucher*innen sozialer Ereignisse bis hin zur Einbeziehung größerer Firmen und Schulen.[5]

Für die Auswahl der E-Mail wurden keinerlei formale oder inhaltliche Präferenzen vorgegeben. Sie erfolgte nach freien Stücken durch die jeweiligen Spender*innen. Obwohl das Korpus auch Threads als Bestandteil einzelner E-Mail-Spenden enthält, handelt es sich insgesamt um ein inkohärentes E-Mail-Korpus.[6] Die zur Spende ausgewählte E-Mail wurde jeweils als Weiterleitungs-E-Mail an die Sammeladresse gesendet und dort gespeichert. Mit dem Empfang einer E-Mail wurde dann ein Fragebogen an die jeweiligen Spender*innen geschickt, mittels dessen demographische Angaben zu den Produzent*innen der E-Mails erhoben werden sollten. Diese Erhebung umfasste neben konventionellen Sozialdaten (Alter, Geschlecht, Ausbildung, Beruf usw.) vor allem auch Angaben zur Häufigkeit der Nutzung von E-Mails im Alltag, zur regionalen Herkunft und sonstigen Rahmenbedingungen der Autor*innen in Hinblick auf ihr Sprachverhalten. Besonders relevant ist in diesem Zusammenhang, dass am Ende des Fragebogens eine juristisch eindeutige Zustimmung der Spender*innen zur Weiterverwendung der zugeschickten E-Mail unter der Voraussetzung, dass deren Identität nicht mehr ersichtlich ist, erbeten wird.[7]

Um eine unrestringierte Weitergabe der E-Mails zu ermöglichen, wurden Wörter und Phrasen, durch die einzelne E-Mail-Verfasser*innen identifiziert werden könnten, in einem nächsten Schritt folglich pseudonymisiert.[8,9] Im Gegensatz zur Anonymisierung wird sensible, d.h. individuenidentifizierende Information bei der Pseudonymisierung

---

[4] Da die Kampagne derzeit noch weitergeführt wird, bitten wir nach wie vor – auch die Leser*innen dieses Beitrags – um E-Mail-Spenden an folgende Adresse: *kodealltag@aau.at*.

[5] Das E-Mail-Segment des Australian National Corpus basiert auf einer anderen Akquisitionsmethode. Dort hatte man anhand eines vordefinierten 8-Kategorienschemas (z.B. Familie, Beschwerden, Liebe/Romantik) einen öffentlichen nationalen Aufruf (Email Australia) veranlasst, der zur Abgabe von über 10.000 E-Mails führte (Lampert 2009).

[6] E-Mail-Korpora gelten dann als formal kohärent, wenn sie sich über ihre Thread-Struktur explizit aufeinander beziehen. Im Falle von CodE Alltag werden die Threads in thread-freie einzelne E-Mails aufgespalten und solche, die die Thread-Struktur bewahren.

[7] Dieser rechtlichen Notwendigkeit sind zahlreiche Spender*innen nicht nachgekommen, so konnte ca. ein Drittel der zugesendeten Datensätze nicht in CodE Alltag aufgenommen werden.

[8] Im Falle des Australian National Corpus zeigte die von den Sender*innen selbst durchgeführte Anonymisierung individueller Daten große Defizite (Lampert 2009), deshalb wurde diese Aufgabe bei CodE Alltag nicht delegiert.

[9] Siehe ähnliche Arbeiten zu klinischen Datensätzen, unter anderem Meystre (2015) oder Stubbs et al. (2015a, 2015b, 2017), aber auch allgemeinere Arbeiten von Medlock (2006).

nicht nur durch Platzhalter wie *XXX* ersetzt, was für viele Anwendungsfälle wie auch für den vorliegenden ungeeignet ist, sondern durch realistische Bezeichnungsalternativen. *Irene Adler* könnte beispielsweise durch *Herta Tschach* ersetzt werden. Die Substitution beschränkt sich jedoch nicht nur auf Vor- und Nachnamen (bei Vornamen werden weibliche und männliche getrennt erfasst), sondern inkludiert auch Namen von Organisationen, Firmen oder Institutionen, Usernamen, Datumsangaben, Orts- und Regionsnamen, Straßennamen, Hausnummern, Postleitzahlen, E-Mail-Adressen, URLs und Domains, Telefon- und Faxnummern sowie Passwörter und IDs jederart als individuenidentifizierende Entitäten-Typen. Bevor spezifische textuelle Erwähnungen (sog. entity mentions) dieser Typen jedoch ersetzt werden können, müssen sie im Text erkannt werden. Mithilfe der Annotationssoftware BRAT[10] (Stenetorp et al. 2012) wurden 1.390 (nach damaligem Stand alle) E-Mails von CodE Alltag$_{S+D}$[11] von drei Annotator*innen und 1.000 E-Mails CodE Alltag$_{XL}$ von fünf Annotator*innen gemäß der verschiedenen Typen manuell annotiert. Diese annotierten E-Mails fungierten als Trainingsdaten für neuronale Machine Learning Modelle zur automatischen Erkennung sensibler Textstellen. Die manuell und maschinell erkannten sensiblen Erwähnungen der oben aufgezählten Entity-Typen wurden im Anschluss daran durch automatisch generierte Alternativen substituiert.[12] Nach der Verschleierung der Identität der E-Mail-Verfasser*innen ist die pseudonymisierte Version von CodE Alltag jetzt für Weiterverwendungen unterschiedlicher Art online zugänglich über https://github.com/codealltag.

Diese beiden Partitionen von CodE Alltag bilden nun den Ausgangspunkt für die weiteren stilometrischen Untersuchungen.

3 Qualitative Stilometrie: Begriffsbestimmung und methodisches Vorgehen

Die Qualitative Stilometrie ist an der Schnittstelle von Linguistik und Computerlinguistik angesiedelt und soll dazu dienen, traditionelle Konzepte aus linguistischen Bereichen wie der Stilistik[13] oder Lexikographie mittels computerlinguistischer Verfahren weiter auszudifferenzieren und zu präzisieren. Dies bildet die Grundlage für die Entwicklung von verschiedenen Werkzeugen, die in Form eines Stilbaukastens für die Analyse digita-

[10] http://brat.nlplab.org/

[11] Bei CodE Alltag$_{S+D}$ wurden zudem Markierungen für unterschiedliche Threads gekennzeichnet, um die Nachrichten später anhand deren zu splitten, sowie Meta-Daten, die anschließend entfernt wurden.

[12] Für eine detaillierte Beschreibung der Erkennungs- und Substitutionsverfahren zur Pseudonymisierung vgl. Eder, Krieg-Holz & Hahn (2019b) und Eder, Krieg-Holz & Hahn (2020).

[13] Vgl. dazu Fleischer, Michel & Starke (1993), Sandig (2006) und Eroms (2008).

ler Editionen und Textkorpora genutzt werden können. Die klassische Stilometrie oder Quantitative Stilistik fokussiert vorrangig Fragestellungen, die die Zuordnung und den Vergleich von Autorenstilen betreffen oder der Identifikation anonymer Autor*innen dienen. Zum stilometrischen Inventar gehören in Bezug auf solche Anwendungen in der Regel formale Verfahren wie Bestimmung von Satzlängen, Wortlängen und lexikalischen Verteilungen (Eder, Kestemont & Rybicki 2016). Gegenüber derartigen Verfahren ist der hier vorgestellte stilometrische Ansatz insofern stärker qualitativ ausgerichtet, als er zum einen das stilometrische Inventar auf prinzipiell alle stilistisch relevanten sprachlichen Kategorien erweitert. Zum anderen ermöglicht er zu zeigen, wie diese stilistischen Kategorien kodiert sind und wie sie miteinander in Verbindung stehen.

Innerhalb der Stilistik ist es unbestritten, dass die Lexik einen sehr direkten Einfluss auf die stilistische Prägung eines Textes hat. Die Untersuchungen zur Qualitativen Stilometrie setzen deshalb im Bereich der Lexik an und gehen grundsätzlich von der Annahme aus, dass es bestimmte Einschätzungsparameter geben muss, bei denen sich z.B. Unterschiede zwischen einem hochsprachlichem und einem vulgären Text signifikant niederschlagen.[14]

Für die differenzierte Wortschatzbeschreibung erfolgt zunächst eine Orientierung an etablierten stilistischen und lexikographischen Kategorien. Dabei wird auf oberster Ebene zwischen Neutralität und Markiertheit unterschieden. Im Falle von Markiertheit[15] kommen Kategorien wie beispielsweise veraltet, gehoben, fachsprachlich, dialektal/regional, derb oder vulgär zum Tragen (Krieg-Holz & Bülow 2016). Derartige Markierungen werden zunächst einzeln ausdifferenziert und sollen dann jeweils über spezielle Lexika abrufbar sein. Später sollen diese Lexika in den Werkzeugkasten integriert werden und ermöglichen, die linguistische Charakteristik von Texten auszumessen.

Den Ausgangspunkt für die differenzierte, stilistische Klassifikation bildet der Bereich der stilschichtig abgesenkten Lexik. Dieser lässt sich stiltheoretisch zunächst von neutralen Elementen unterscheiden. Von (virtueller) stilistischer Neutralität (Eroms 2008)

---

[14] Beispiel für vulgären Sprachgebrauch finden sich etwa im Cybermobbing-Korpus von Marx (2017, 186), z.B. „ba alter die geht auf babystrich (isg, pg_1_rüd_stu, 2011-04-14, 12:50:46)".

[15] Das Konzept der Markiertheit geht ursprünglich auf die Natürlichkeitstheorie zurück, einem linguistischen Ansatz, der die Sprecher*innen mit ihren produktiven und rezeptiven Fähigkeiten in den Vordergrund stellt. Anwendung fand dies vor allem in der Phonologie und Morphologie (vgl. Mayerthaler 1981, Wurzel 1984), wobei es im Wesentlichen darum ging, was für den Sprecher besser oder schlechter ist, was weniger oder mehr markiert ist. Eroms (2008, 60f.) überträgt die Differenzierung zwischen Markiertheit und Neutralität auf die Stilistik und unterscheidet zwei Faktoren, die zu virtueller stilistischer Neutralität führen können: Systemzwang und Systemneutralisierung. Ersteres liegt im Falle eines einelementigen Paradigmas vor, in dem es keine Varianten gibt, also überhaupt nur ein Ausdruck zur Verfügung steht. Um Systemneutralisierung handelt es sich, wenn innerhalb eines Wortfeldes (im Sinne einer Synonymengruppe) ein Ausdruck ein geringeres stilistisches Potential hat, weil er in allen Kommunikationsbereichen angemessen ist.

wird dann gesprochen, wenn ein Lexem ohne Einschränkung in allen Kommunikationsbereichen und Textsorten vorkommen kann ohne eine besondere Wirkung zu erzielen. Auch innerhalb der Lexikographie geht man i.d.R. von einem unmarkierten Zentrum und einer markierten Peripherie aus. Dementsprechend werden Lemmata mit speziellen Markierungen bzw. diasystematischen Angaben versehen. Eine Bedeutung für den abgesenkten Bereich haben dabei ganz verschiedene Kriterien oder Mikrosysteme. Hierzu gehören die Markierungsarten *diamedial* (z.B. umgangssprachlich), *diastratisch* (z.B. Slang), *diaphasisch* (z.B. informell), *diatopisch* (z.B. bairisch) und *diaevaluativ* (z.B. pejorativ). Sie werden nicht durchgehend übereinstimmend angewendet, sondern gehen vielfach ineinander über und wirken unscharf. In der aktuellen Diskussion um Cyber-Mobbing und Hassrede kommen weitere Benennungen dazu, die versuchen, diverse Abstufungen von Beleidigungen, Beschimpfungen abzubilden. Sie fallen primär in den diaevaluativen Bereich. Eine Skala, die in diesem Zusammenhang in computerlinguistischen Untersuchungen angewendet wurde, unterscheidet zwischen *profanity – insult – abuse* (Wiegand et al. 2018).

Word Embeddings etablieren sich zunehmend als neuartige korpuslinguistische Methode auf der Grundlage der distributionellen Semantik. Sie werden im Folgenden für die automatische Stilbeschreibung genutzt, wobei sie zum einen dazu dienen sollen, die stilistische Qualität von Texten auszumessen, und so etwa in Bezug auf Dichotomien wie *formell – informell*, *vulgärsprachig – hochsprachig* valide Aussagen darüber zu ermöglichen, worin sich die stilistische Gestaltung dieser Texte unterscheidet. Somit könnten auch Graduierungen vorgenommen werden (sehr vulgär, weniger vulgär usw.). Zum anderen verwenden wir Word Embeddings für den Lexikonaufbau. Mit Hilfe von Word Embeddings können die lexikalisch-semantische Ähnlichkeiten zwischen Lexemen berechnet werden. Grundlage hierfür sind Vektorrepräsentationen, die sich aus dem Kotext von Wörtern in Sätzen ableiten. Damit können u.a. Daten zu kookkurrierenden Elementen generiert und Analogieschlüsse gezogen werden (Mikolov et al. 2013).

Um die Subjektivität der Annotationen (z.B. Abstufungen zwischen vulgärsprachlich und standardsprachlich) zu verringern, integriert die Methode der Qualitativen Stilometrie Crowdsourcing-Verfahren zur Einschätzung semantischer Unterscheidungen. Diese nutzen eine möglichst heterogene Menge von unbekannten Akteur*innen, um Aufgaben, die traditionell intern von einigen wenigen Personen durchgeführt werden, zu übernehmen und auf eine breitere quantitative Basis zu stellen. Über das Internet und diverse Crowdsourcing-Anbieter lässt sich durch derartige Auslagerungsstrategien die Zahl an Einschätzer*innen beträchtlich erhöhen.

4 Lexikon für stilistisch abgesenkte Sprache: VulGer

Eine zentrale Grundlage für den Stilwerkzeugkasten sollen Lexika bilden, in denen bestimmte Bereiche stilistisch relevanter Markierungen so präzise wie möglich abgebildet werden. In einem ersten Schritt wurde dazu zunächst ein Lexikon für den abgesenkten Bereich aufgebaut (VulGer, vgl. Eder, Krieg-Holz & Hahn (2019a)). Der Prozess zur Erstellung orientiert sich an der Arbeit von Wiegand et al. (2018), wobei dieses Lexikon anders als herkömmliche Ansätze, die Lexeme in Kategorien (zum Beispiel binär abgesenkt versus nicht abgesenkt oder abwertend, derb, vulgär, etc.) einteilen, lexikalische Elemente auf einer Skala von neutral bis vulgär graduieren soll. Zudem erfolgte die Annotation der Grade nicht durch einzelne Lexikograph*innen, sondern wurde mittels Crowdsourcing durchgeführt.

4.1 Aufbau

Zuerst wurde für den Aufbau des Lexikons auf schon vorhandene lexikalische Ressourcen zurückgegriffen. Aus dem deutschen Wiktionary und dem deutschen Online-Wörterbuch OpenThesaurus wurden alle Einträge, die als vulgär, derb oder abwertend[16] klassifiziert waren, extrahiert, wobei vorerst lediglich Einzellexeme Berücksichtigung fanden, da klassische Word Embeddings einzelwortbasiert sind. Vorhandene Wortbildungselemente mit klarem Bezug zu abgesenkter Sprache, wie *geil*, *scheiß*, *drecks*, deren Vulgarität je nach Komposition variieren kann, wurden dementsprechend durch Komposita mit diesen Elementen, die in den im Folgenden beschriebenen Korpora beziehungsweise Word Embeddings enthalten waren, ersetzt. Die daraus resultierende Liste abgesenkter Lexeme wurde um distributionell ähnliche Wörter erweitert (siehe auch Tulkens et al. (2016), Wiegand et al. (2018)). Dazu wurden zum einen FastText-Word Embeddings (Grave et al. 2018) verwendet, zum anderen wurden aus CodE Alltag und dem Dortmunder Chat Korpus (Beißwenger 2013) Word2Vec-Embeddings (Mikolov et al. 2013) erzeugt. Für letzteres wurde das Gensim-Modul (Řehůřek & Sojka 2010) genutzt, das auch zur Berechnung der semantisch benachbarten Wörter auf Basis der genannten Word Embeddings eingesetzt wurde. Die mit diesem Vorgehen einhergehenden flektierten Formen und Falschschreibungen wurden manuell bearbeitet und gegebenenfalls beseitigt. Hingegen wurden berechnete semantische Nachbarwörter, die mutmaßlich neutral sind, nicht aus der Liste entfernt, weil sie für unseren Anwendungsfall der Graduierung von Vulgarität von Nutzen sind.

Insgesamt umfasst das in diesem Schritt entstandene Basislexikon für die weitere Annotation 3.300 Einträge.

[16] Inklusive der Abkürzungen vulg., vul. und abw.

4.2 Annotation

Die Annotation des Lexikons wurde neben der Crowdsourcing-Plattform FIGURE EIGHT[17] vor allem mit dem deutschen Crowdsourcing-Anbieter CLICKWORKER[18] umgesetzt. Für die Skalierung der Einträge selbst kam Best-Worst-Scaling (BWS) zum Einsatz. Diese von Louviere, Flynn & Marley (2015) entwickelte Methode verspricht Annotationen von hoher Qualität, wie Kiritchenko & Mohammad (2017) zeigen, die BWS zur Bewertung emotionaler Sprache eingesetzt haben (Kiritschenko & Mohammad 2016, 2017). Beim BWS werden den Annotator*innen jeweils *n* Begriffe vorgelegt, wobei *n* üblicherweise 4 ist. Dann müssen sie aus diesem *n*-Tupel den besten Begriff, der das gewählte Kriterium am besten erfüllt, und den schlechtesten Begriff, auf den die Eigenschaft folglich am wenigsten zutrifft, wählen. In dem hier beschriebenen konkreten Fall wurde dementsprechend nach dem neutralsten und nach dem vulgärsten Wort in einem 4er-Tupel gefragt. Dabei wurden die Einschätzer*innen mit folgendem Text an ihre Aufgabe herangeführt:

Wörter können neutral sein, das heißt sie können überall (Nachrichten etc.) verwendet werden, z.B. *Schmetterling*. Sie können aber auch auf die Umgangssprache beschränkt (z.B. *kotzen*) oder derb (z.B. *verrecken*) und schon richtig vulgär (z.B. *Pissfotze*) sein. Dabei sind die Übergänge fließend. In diesem Job sollen Sie bestimmen, wie vulgär Wörter sind.
Ihnen werden vier Begriffe präsentiert. Zu diesen bekommen Sie jeweils zwei Aufgaben beziehungsweise Fragen:

1. Neutralstes Wort
   Erstens sollen Sie den Begriff wählen, der von diesen vier Wörtern am neutralsten ist, also auf einer Skala von neutral über umgangssprachlich, abwertend und derb bis hin zu vulgär am ehesten in Richtung neutral geht.
2. Vulgärstes Wort
   Zweitens sollen Sie den Begriff wählen, der von diesen vier Wörtern am vulgärsten ist, also auf der Skala am nächsten bei vulgär zu finden ist.

---

[17] https://www.figure-eight.com/

[18] https://www.clickworker.de/

Eine Beispielaufgabe, wie sie bei Clickworker aussehen könnte, zeigt folgendes Bild.

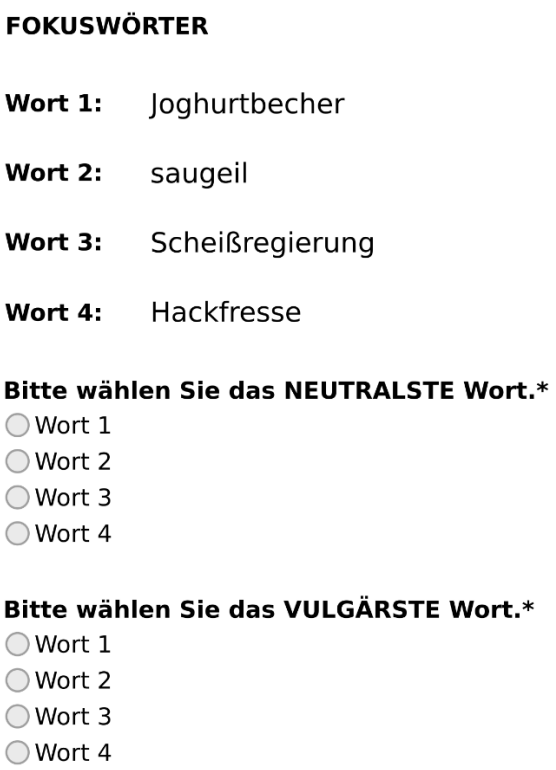

Abbildung 1: Beispielaufgabe mit 4er-Tupel auf der Crowdsourcing-Plattform Clickworker

Um aus den 3.300 Einträgen im Basislexikon 6.600 4er-Tupel zu generieren, wurde das BWS-Tool[19] von Kiritchenko & Mohammad (2016, 2017) verwendet, das die Wort-Tupel willkürlich zusammenstellt unter der Voraussetzung, dass eine Tupelzusammensetzung nur einmal vorkommt und ein Wort nicht mehr als einmal in acht verschiedenen Tupeln auftritt. Jedes der erzeugten Tupel wurde fünf Mal von verschiedenen Crowdworkern bearbeitet. Insgesamt wurden die 3.300 Basiswörter folglich mit 33.000 Bewertungen versehen. Ebenfalls auf ein Programm von Kiritchenko & Mohammad (2016, 2017) zurückgreifend wurde anschließend Counts Analysis (Orme 2009) angewendet, um aus der Gesamtheit der Bewertungen einzelne Scores zu berechnen. Dabei wird der prozentuale Anteil von Bewertungen, in denen ein Wort als schlechtestes (d. h. vulgärstes) gewählt wurde, von dem der Bewertungen als bestes Wort subtrahiert, was Scores zwischen +1 (am neutralsten) und -1 (am vulgärsten) ergibt.

[19] http://www.saifmohammad.com/WebPages/BestWorst.html

4.3 Erweiterung

An Arbeiten zur Emotionserkennung von Lexemen (Li et al. 2017, Buechel & Hahn 2018) orientiert, wurde das Lexikon in einem nächsten Schritt automatisch erweitert, indem mithilfe von Verfahren des Maschinellen Lernens Scores für noch unberücksichtigte, neue Wörter berechnet wurden. Das Basislexikon fungierte dabei als Trainingsgrundlage für ein Ridge Regression-Modell.[20] Dieses lernt aus Word Embeddings, die als Features für die einzelnen Einträge im Lexikon dienten, die passenden Scores.[21] Als Word Embeddings wurden FastText-Embeddings (Grave et al. 2018) verwendet, wobei dort nicht vorhandene Wörter unberücksichtigt blieben. Nach dem Training wurde das Modell auf neue Wörter angewendet, die thematisch in Frage kommen. Schimpfwörter weisen offenkundig Überschneidungen mit abgesenkten Lexemen auf. Zwar sind nicht alle vulgären Wörter auch Schimpfwörter (bspw. *pissen*) und genauso haben nicht alle Beschimpfungen denselben Grad an Vulgarität (bspw. *Blödmann*, *Jasager* oder *Clown*), dennoch hat der hier verwendete Ansatz der Skalierung genau hier seine Stärke und kann den unterschiedlichen Vulgaritätsgraden gerecht werden.

Zunächst wurden Scores für drei deutsche Schimpfwortlisten errechnet: Hyperhero[22], insult.wiki[23] und schimpfwoerter.de[24], wobei Lexeme, die im Basislexikon schon vorhanden waren, ausgeklammert wurden. Gleiches gilt für Einträge, für die auch ohne deren Groß- und Kleinschreibung zu beachten kein entsprechendes FastText-Word Embedding gefunden wurde, wodurch sehr seltene Wörter ebenfalls exkludiert werden konnten. Auf diesem Weg konnte das Lexikon um 2.046 Wörter erweitert werden.

Außerdem wurde Gebrauch von Korpora gemacht, die als Trainingsmaterial für die Erkennung von Hasssprache in deutschen Tweets dienen, da angenommen werden kann, dass auch sie tendenziell einen abgesenkteren Sprachgebrauch aufweisen. IWG Hatespeech[25] (Ross et al. 2016) enthält ungefähr 500 deutsche Tweets, während das Korpus zur GermEval 2018 Challenge[26] (Wiegand et al. 2018b) mit mehr als 8.500 Tweets wesentlich größer ist. Die Tweets beider Korpora wurden von menschlichen Annotator*innen neben einer differenzierteren Klassifizierung binär als Hasssprache und keine

[20] Dazu wurde die Implementierung von sklearn.org mit den Standardparametern verwendet.

[21] Weitere Informationen und eine Evaluation zu unterschiedlichen Regressionsmodellen und Word Embeddings finden sich in Eder, Krieg-Holz & Hahn (2019a).

[22] http://www.hyperhero.com/de/insults.htm

[23] http://www.insult.wiki/wiki/Schimpfwort-Liste

[24] https://www.schimpfwoerter.de/

[25] https://github.com/UCSM-DUE/IWG_hatespeech_public

[26] https://projects.cai.fbi.h-da.de/iggsa/

Hasssprache bewertet.[27] Aus allen Tweets, die mindestens einmal als Hasssprache kategorisiert wurden, wurden Wörter extrahiert, wobei Stoppwörter, Hashtags und Token, die kürzer als 4 Buchstaben sind oder nicht-alphabetischen Zeichen mit Ausnahme von Bindestrichen enthalten, keine Berücksichtigung fanden. Die ausgelesenen Wörter wurden anschließend mit SPACY[28] (Honnibal & Montani 2017) weitestmöglich lemmatisiert und bezüglich Groß- und Kleinschreibung normalisiert. Auch hier wurden Einträge ohne entsprechendes Word Embedding exkludiert, um nicht nur seltene, sondern gegebenenfalls auch orthographisch inkorrekte Token zu entfernen. Trotz dieser Bereinigungsstrategien ist die auf diese Art und Weise akquirierte Wortliste mit 5.700 Einträgen stärker verrauscht als die Lexeme, die aus lexikalischen Ressourcen wie Schimpfwortlisten extrahiert wurden.

Mit 3.300 Lexemen aus dem Basislexikon, 2.046 Schimpfwörtern und den 5.700 Wörtern aus Tweets mit Hasssprache umfasst das finale erweiterte VULGER-Lexikon nun insgesamt 11.046 Einträge.

Mit dem vorliegenden Beitrag sollte gezeigt werden, auf welche Weise das vulgärsprachliche Lexikon VULGER aufgebaut wurde. Es stellt einen wichtigen Schritt für die Entwicklung eines Stilwerkzeugkastens dar, der zur automatischen Erkennung und Klassifikation stilistischer Merkmale genutzt werden soll und sich gegenüber bereits existierenden Methoden der klassischen Stilometrie stärker an der Qualität der stilistischen Einzelmerkmale orientiert, um (semi-)automatische Stilanalysen im Sinne einer „Qualitativen Stilometrie“ zu ermöglichen. Im Mittelpunkt der Ausführungen stand die Erläuterung der notwendigen infrastrukturellen Voraussetzungen, wie auch der Akquise eines stilistisch breitbandig angelegten E-Mail-Korpus. Zunächst auf Basis von CODE ALLTAG sollen mithilfe des Stilbaukastens und VULGER Analysen dieses stilistischen Variantenreichtums ermöglicht werden.

Literatur

Becker, Markus, Andrew Bredenkamp, Berthold Crysmann und Judith Klein. 2003. „Annotation of error types for German newsgroup corpus.“ In *Treebanks. Building and Using Parsed Corpora*, hrsg. v. Anne Abeillé, 89–100. Dordrecht: Springer Netherlands.

Beißwenger, Michael und Lothar Lemnitzer. 2013. „Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt ‚Digitales Wörterbuch der deutschen Sprache‘ (DWDS)“. In *Journal for Language Technology and Computational Linguistics* 28 (2): 1–22.

[27] Für IWG HATESPEECH: HatespeechOrNot? Yes oder No; für GERMEVAL 2018: Offense oder Other.

[28] https://spacy.io/

Beißwenger, Michael. 2013. „Das Dortmunder ChatKorpus." In *Zeitschrift für germanistische Linguistik*, 41(1): 161–164.

Buechel, Sven und Udo Hahn. 2018. „Word emotion induction for multiple languages as a deep multi-task learning problem." In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018, volume 1: Long Papers*, 1907–1918. Stroudsburg, PA. Association for Computational Linguistics (ACL).

Cormack, Gordon V. 2007. „TREC 2007 Spam Track overview. In *TREC 2007 – Proceedings of the 15th Text REtrieval Conference. Gaithersburg, Maryland, November 5–9, 2007*.

Declerck, Thierry und Judith Klein. 1997. „Ein Email-Korpus zur Entwicklung und Evaluierung der Analysekomponente eines Terminvereinbarungssystems." In *DGfS-CL '97 – Proceedings der 6. Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft: Integrative Ansätze in der Computerlinguistik. Heidelberg, Deutschland, 8.–10. Oktober 1997*.

Eder, Elisabeth, Ulrike Krieg-Holz und Udo Hahn. 2019a. „At the Lower End of Language – Exploring the Vulgar and Obscene Side of German." In *Proceedings of the Third Workshop on Abusive Language Online. Florence, Italy, August 1, 2019*, 119–128. Stroudsburg, PA. Association for Computational Linguistics (ACL).

Eder, Elisabeth, Ulrike Krieg-Holz und Udo Hahn. 2019b. „De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus." In *Proceedings of Recent Advances in Natural Language Processing (RANLP). Varna, Bulgaria, 2–4 September 2019*. 259–269.

Eder, Elisabeth, Ulrike Krieg-Holz und Udo Hahn. 2020. „CodE Alltag 2.0 – A Pseudonymized German-Language Email Corpus." In *Proceedings of the International Conference on Language Ressources and Evaluation (LREC). Marseille, France, 13–15 May 2020*. 4468–4479.

Eder, Maciej, Mike Kestemont und Jan Rybicki. 2016. „Stylometry with R: a package for computational text analysis." In *R Journal*, 16 (1):107-121.

Eroms, Hans-Werner. 2008. *Stil und Stilistik. Eine Einführung*. Berlin: Erich Schmidt.

Fleischer, Wolfgang, Michel, Georg und Starke, Günter. 1993. *Stilistik der deutschen Gegenwartssprache*. Frankfurt a. M. u.a.: Lang.

Geyken, Alexander. 2007. „The DWDS corpus: a reference corpus for the German language of the 20th century." In *Collocations and Idioms: Corpus-based Linguistic and Lexicographic Studies*, hrsg. v.Christiane Fellbaum, 23–40. London: Continuum.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin und Tomáš Mikolov. 2018. „Learning word vectors for 157 languages." In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, 3483–3487. Paris: European Language Resources Association (ELRA).

Kiritchenko, Svetlana und Saif M. Mohammad. 2016. „Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling." In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12-17, 2016*, 811–817. Stroudsburg, PA: Association for Computational Linguistics (ACL).

Kiritchenko, Svetlana und Saif M. Mohammad. 2017. „Best-worst scaling more reliable than rating scales: a case study on sentiment intensity annotation." In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, British Columbia, Canada, July 30-August 4, 2017, volume 2: Short Papers*, 465–470. Stroudsburg, PA: Association for Computational Linguistics (ACL).

Klimt, Bryan und Yiming Yang. 2004. „The Enron corpus: a new dataset for email classification research." In *ECML 2004 – Proceedings of the 15th European Conference on Machine Learning. Pisa, Italien, 20.–24. September 2004*, 217–226 (LNCS, 3201). Berlin, Heidelberg: Springer.

Krieg-Holz, Ulrike, Christian Schuschnig, Franz Matthies, Benjamin Redling und Udo Hahn. 2016. „CODE ALLTAG: A German-language e-mail corpus." In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, 2543–2550. Paris: European Language Resources Association (ELRA-ELDA).

Krieg-Holz, Ulrike und Lars Bülow. 2016. *Linguistische Stil- und Textanalyse. Eine Einführung*. Tübingen: Narr/Francke/Attempto.

Krieg-Holz, Ulrike und Udo Hahn. 2016. „CODE ALLTAG: Ein deutsches E-Mail-Korpus für die Forensische Linguistik." In Performativität in Sprache und Recht, hrsg. v. Lars Bülow, Jochen Bung, Rüdiger Harnisch und Rainer Wernsmann, 245–264. Berlin, Boston: de Gruyter.

Kupietz, Marc und Harald Lüngen. 2014. „Recent developments in DeReKo." In *LREC 2014 – Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Island, 26.–31. Mai 2014*, 2378–2385.

Lampert, Andrew. 2009. „Email in the Australian National Corpus." In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*. 55–60. Somerville, MA: Cascadilla Proceedings Project.

Louviere, Jordan J., Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge: Cambridge University Press.

Marx, Konstanze. 2017. *Diskursphänomen Cybermobbing. Ein internetlinguistischer Zugang zu [digitaler] Gewalt*. (Diskursmuster – Discourse Patterns 17). Berlin, Boston: De Gruyter.

Mayerthaler, Willy. 1981. *Morphologische Natürlichkeit*. Wiesbaden: Akademische Verlagsgesellschaft Athenaion.

Medlock, Ben. 2006. „An introduction to NLP-based textual anonymisation." In *LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation. Genua, Italien, 22.–28. Mai 2006*, 1051–1056.

Meystre, Stéphane M. 2015. „De-identification of unstructured clinical data for patient privacy protection." In *Medical Data Privacy Handbook*, hrsg. v. Aris Gkoulalas-Divanis und Grigorios Loukides, 697–716. Cham, Heidelberg, New York, Dordrecht, London: Springer International Publishing.

Mikolov, Tomáš, Ilya Sutskever, Kai Chen, Gregory S. Corrado und Jeffrey Dean. 2013. „Distributed representations of words and phrases and their compositionality." In *Advances in Neural Information Processing Systems 26 — NIPS 2013. Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA, December 5-10, 2013*, 3111–3119. Red Hook, NY: Curran Associates, Inc.

Orme, Bryan. 2009. „Maxdiff analysis: simple counting, individual-level logit, and HB." *Sawtooth Software*, Inc.

Řehůřek, Radim und Petr Sojka. 2010. „Software framework for topic modelling with large corpora." In *Proceedings of the Workshop on New Challenges for NLP Frameworks @ LREC 2010. La Valletta, Malta, May 22, 2010*, 45–50. Paris: European Language Resources Association (ELRA).

Sandig, Barbara. 2006. *Textstilistik des Deutschen*. Berlin, New York: de Gruyter.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou und Jun'ichi Tsujii. 2012. „BRAT: a Web-based tool for NLP-assisted text annotation." In *EACL 2012 — Proc. of the 13th Conf. of the European Chapter of the Association for Computational Linguistics: Demonstrations. Avignon, France, April 25-26, 2012*, 102–107.

Storrer, Angelika. 2013. „Sprachstil und Sprachvariation in sozialen Netzwerken." In *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*, hrsg. v. Barbara Frank-Job, Alexander Mehler und Tilmann Sutter, 329–364. Wiesbaden: VS Verlag für Sozialwissenschaften.

Stubbs, Amber und Özlem Uzuner. 2015a. „Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus." In *Journal of Biomedical Informatics* 58 (Supplement): 20–29.

Stubbs, Amber, Özlem Uzuner, Christopher Kotfila, Ira, Goldstein und Peter Szolovits. 2015b. „Challenges in synthesizing surrogate PHI in narrative EMRs." In *Medical Data Privacy Handbook, hrsg. v. Aris Gkoulalas-Divanis und Grigorios Loukides*, 717–735. Cham, Heidelberg, New York, Dordrecht, London: Springer International Publishing.

Stubbs, Amber, Michele Filannino und Özlem Uzuner. 2017. „De-identification of psychiatric intake records: overview of 2016 CEGS NGRID Shared Tasks Track 1." In *Journal of Biomedical Informatics* 75 (Supplement): 4–18.

Tulkens, Stéphan, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven und Walter Daelemans. 2016. „A dictionary-based approach to racism detection in Dutch social media." In *TA-COS 2016 — Proceedings of the Workshop on Text Analytics for Cybersecurity and Online Safety @ LREC 2016. Portorož, Slovenia, 23 May 2016*, 11–17.

Wiegand, Michael, Josef Ruppenhofer, Anna Schmidt und Clayton Greenberg. 2018. „Inducing a lexicon of abusive words: a feature-based approach." In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018, volume 1: Long Papers*, 1046–1056. Stroudsburg, PA: Association for Computational Linguistics (ACL).

Wurzel, Wolfgang U. 1984. *Flexionsmorphologie und Natürlichkeit*. Studia Grammatica XXI. Berlin: Akademie-Verlag.