Parameter zur Klassifizierung stilistischer Varianz bei E-Mails

Ulrike Krieg-Holz

Institut für Germanistik
Universität Leipzig

ulrike.krieg-holz@uni-leipzig.de

Udo Hahn

Institut für Germanistische Sprachwissenschaft
Friedrich Schiller-Universität Jena

udo.hahn@uni-jena.de

Zusammenfassung. Auf der empirischen Grundlage eines im Aufbau befindlichen deutschsprachigen E-Mail-Korpus werden Aspekte der stilistischen Varianz innerhalb der Kommunikationsform ‚E-Mail' untersucht.

Mit der wachsenden Bedeutung der empirischen Fundierung linguistischer Analysen ist auch für die deutsche Sprache eine große Vielfalt von Korpora entstanden.[1] Aus linguistischer Sicht stand dabei lange die Idee der Erfassung möglichst vielfältiger und großvolumiger Mengen von sprachlichen Rohdaten im Vordergrund,[2] die auch dem Anspruch genügen sollten, den Sprachgebrauch weitgehend „repräsentativ" abzubilden (exemplarisch gilt dies etwa für das DeReKo-Korpus[3] [KBKW10] sowie das DWDS-Kernkorpus[4] [Geyk07]). In diesen Korpora spielen nach wie vor Zeitungstexte die weithin dominierende Rolle, ergänzt durch überwiegend literarisch-belletristische Quellen und eher geringe Vorkommen von Gebrauchstexten (Kochrezepte, Montageanleitungen usw.). Trotz der angestrebten Vielfalt von Textsorten liegt der Schwerpunkt der in diesen Korpora auftretenden Texte eindeutig im Bereich der deutschen Standardsprache mit klarer Ausrichtung auf formelle Kommunikation.

Auf dem entgegengesetzten Pol des Kontinuums formeller vs. informeller Sprachgebrauch können aktuelle Arbeiten zur Erhebung von Korpora geschriebener Alltagssprache eingeordnet werden [Stor13]. Die Sammlung und Annotation informeller Sprache ist derzeit für das Deutsche am weitesten bei DeRiK gediehen [BEG+13], einem Korpus zur Erfassung computervermittelter Kommunikation (Blogs, Chats usw.) als Ergänzung des DWDS-Kernkorpus. In dieser Textkollektion werden jedoch explizit keine E-Mails berücksichtigt.

Wegen ihrer bedeutenden Rolle im öffentlichen wie privaten Kommunikationsumfeld moderner IT-basierter Gesellschaften ist damit ein bedeutsames korpuslinguistisches Desiderat beschrieben – reichen doch die Textsorten der Kommunikationsform ‚E-Mail' inhaltlich mittlerweile von zum Teil hochgradig formalisierten professionellen Diskursen (Geschäfts- bzw. Verwaltungspost) bis hin zu gänzlich persönlichen und somit rein informellen Interaktionen. Deshalb werden sie in Bezug auf ihre

[1] Eine aktuelle Übersicht enthält http://de.clarin.eu/de/sprachressourcen/corpora.html (letzter Aufruf: 3.11.2014)
[2] Die sprachlichen Rohdaten sind zum Teil bereits auch automatisch lemmatisiert und nach Wortarten annotiert (POS-Tagging).
[3] http://www1.ids-mannheim.de/kl/projekte/korpora/ (letzter Aufruf: 3.11.2014)
[4] http://www.dwds.de/ressourcen/korpora/ (letzter Aufruf: 3.11.2014)

stilistische Ausformung äußerst unterschiedlich gestaltet und bilden damit – so unsere leitende Hypothese – innerhalb einer Kommunikationsform eine sehr große Bandbreite der performativen Varianz auf dem gesamten Kontinuum formeller und informeller Sprache ab. E-Mails eignen sich somit ganz besonders für empirisch fundierte stilistische Untersuchungen.

Im Vergleich zu literaturwissenschaftlichen Stilanalysen, in deren Mittelpunkt die stilistische Figuriertheit von Texten oder auch Autorenstile stehen, fokussiert der sprachwissenschaftliche Stilbegriff ganz allgemein die Spezifik der sprachlichen Ausgestaltung von Textstrukturen. Diese Spezifik sprachlicher Formulierungen resultiert prinzipiell 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 ganz entscheidend prägen. Zum einen können sprachliche Handlungen desselben Typs auf verschiedene Weise durchgeführt werden, zum anderen unterscheiden sich sprachliche Handlungen verschiedenen Typs in stilistischer Hinsicht. Dadurch können sowohl Textproduktions- als auch Textrezeptionsprozesse erheblich beeinflusst werden, denn das Wissen über Stil ist Teil der Textsortenkompetenz, also der Fähigkeit, auf der Grundlage eines mehr oder weniger bewussten Wissens über Textsortenqualitäten in der Kommunikation operieren zu können.

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; es gilt also, die Träger stilistischer Information zu charakterisieren. Dazu hat sich ein terminologisches Inventar herausgebildet, das zum Teil durch verschiedene Beschreibungsansätze geprägt und entsprechend ungleich stark etabliert ist. So bezeichnen die Begriffe ‚Stilelement' und ‚Stilzug' ursprünglich grundlegende funktionalstilistische Kategorien (vgl. Fleischer et al. 1993 [FlMS93, S. 27]), wobei sich Stilelemente immer auf einzelne sprachliche Mittel innerhalb eines Relationsgefüges beziehen (z.B. markierte lexikalische Elemente, eine spezifische Wortstellung oder bestimmte Elemente auf der lautlichen und graphemisch-ikonischen Ebene). Demgegenüber werden als Stilzüge bestimmte Stilstrukturen zusammengefasst, die aus einer typischen Kombination verschiedener Stilelemente resultieren. Sie lassen sich als Bündel miteinander vorkommender, kookkuierender Merkmale auffassen, die als eine bedeutsame, sinnhafte Gestalt interpretiert werden können (vgl. Selting & Hinnenkamp 1989 [SeHi89, S.5f.]; Sandig 2006 [Sand06, S. 54f.]). Für die Beschreibung derartiger Merkmalbündel bzw. Stilzüge steht eine breite, bisher systematisch kaum erfasste Vielfalt an Kriterien zur Verfügung, aus denen hier diejenigen gefiltert werden sollen, die innerhalb der Textsorten der E-Mail-Kommunikation distinktiv wirken.

Um unsere Arbeiten auf eine solide empirische Basis zu stellen, wird derzeit am Institut für Germanistik der Universität Leipzig und am Institut für Germanistische Sprachwissenschaft der Friedrich Schiller-Universität Jena am Aufbau eines umfassenden **Ko**rpus **d**eutschsprachiger **E**mails (KodE Alltag) gearbeitet. Alle bislang verfügbaren E-Mail-Korpora (hier ist vor allem das Enron-Korpus [KlYa04] zu erwähnen) erfassen nur die englische Sprache, für das Deutsche existiert bislang kein vergleichbares Korpus. Auch diese Lücke wollen wir mit unseren Arbeiten füllen. In unserem Beitrag werden wir die leitenden Entwurfsprinzipien für und den konkreten Aufbau von KodE Alltag sowie den aktuellen

Stand der Datenerhebung im Detail beschreiben. Hierzu gehören auch Ausführungen zu Aspekten des Urheberrechts an E-Mails und zu ihrer (semi-automatischen) Anonymisierung.

Im Zentrum unseres Beitrags werden jedoch erste Befunde zur Klassifizierung der stilistischen Varianz in KodE Alltag stehen. In diesem Zusammenhang unterscheiden wir formellen vs. informellen sprachlichen Stil graduell anhand einer Fülle von Parametern, die jeweils in Verbindung mit anderen zur Ausprägung entsprechender Stilzüge bzw. Merkmalbündel beitragen können. Dazu gehören

- die lexikalische Vielfalt (in Bezug auf kanonische Lexika) und Variabilität (in Bezug auf alternative Korpora wie DeReKo/DWDS oder DeRiK),
- die Orientierung an Mündlichkeit (z.B. klitisierte oder reduzierte Wortformen, spontansprachliche Syntax),
- die Distribution von Abkürzungen und Icons,
- Abweichungen in Formenbildung, Orthographie und Interpunktion usw.

Anhand solcher Parameter wird eine Untergliederung von KodE Alltag in Subkorpora möglich sein, die unterschiedliche Grade an Formalisierung von Sprache in Form von Stilzügen ausdrücken. Diese Stilzüge werden dann als Grundlage für das Training von statistischen Klassifikatoren genutzt, um ungesehene Sprachdaten entlang unterschiedlicher Stilformen automatisch klassifizieren zu können.

Unsere Arbeiten verbinden also germanistische Stilforschung und Korpuslinguistik mit Verfahren der automatischen Textklassifikation aus dem Bereich der Computerlinguistik. Die von uns erarbeiteten Ergebnisse werden wir im Rahmen von Fragestellungen aus dem Bereich der Forensischen Linguistik (Autorenerkennung durch stilistische Write-Prints) auf Anwendbarkeit prüfen.

Literatur

[BEG+13] Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer & Angelika Storrer (2013). DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531-537.

[FlMS93] Wolfgang Fleischer, Georg Michel & Günter Starke (1993). *Stilistik der deutschen Gegenwartssprache.* Tübingen: Niemeyer.

[Geyk07] Alexander Geyken (2007). The DWDS corpus: a reference corpus for the German language of the 20th century, In: Christiane Fellbaum (Ed.), *Collocations and Idioms*. London: Continuum, pp. 23–40.

[KBKW10] Marc Kupietz, Cyril Belica, Holger Keibel & Andreas Witt (2010). The German reference corpus DeReKo: a primordial sample for linguistic research. In: *LREC '10 – Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta, May 2010, pp. 1848-1854.

[KlYa04] Bryan Klimt & Yiming Yang (2004). The Enron corpus: a new dataset for email classification research. In: Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti & Dino Pedreschi (Eds.), *ECML 2004 - Proceedings of the 15th European Conference on Machine Learning*, Pisa, Italy, September 20-24, 2004. Berlin, Heidelberg: Springer, pp.217–226 (Lecture Notes in Computer Science, 3201)

[Sand06] Barbara Sandig (2006). *Textstilistik des Deutschen*. Berlin, New York: de Gruyter.

[SeHi89] Margret Selting & Volker Hinnenkamp (1989). Stil und Stilisierung in der interpretativen Soziolinguistik. In: Volker Hinnenkamp & Margret Selting (Eds.), *Stil und Stilisierung*. Tübingen: Niemeyer, pp. 1-23.

[Stor13] Angelika Storrer (2013). Sprachstil und Sprachvariation in sozialen Netzwerken, in: Barbara Frank-Job, Alexander Mehler & Tilmann Sutter (Eds.), *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*, Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 329–364.