Sprachliche Variation in der Germanistik: eine n-Gramm-basierte Stilanalyse

Andresen, Melanie
Melanie.Andresen@uni-hamburg.de
Universität Hamburg, Deutschland

Einleitung

An zahlreichen Universitäten werden die wissenschaftlichen Disziplinen Linguistik und Literaturwissenschaft in einem gemeinsamen Studiengang angeboten, der beispielsweise „Germanistik“ oder „Deutsche Sprache und Literatur“ heißt. Dies suggeriert eine große fachliche Nähe dieser Disziplinen, die jedoch im Selbstverständnis der meisten Wissenschaftler/innen dieser Fächer keine Entsprechung hat. Linguistik und Literaturwissenschaft unterscheiden sich in ihrem Erkenntnisinteresse, ihren Methoden und auch in ihrer Sprache, wie punktuell bereits beschrieben wurde: So stellt Haggan (2004) bei der Untersuchung von Titeln wissenschaftlicher Publikationen in Linguistik, Literatur- und Naturwissenschaft fest, dass die Sprache der Literaturwissenschaft sich (auch) an ästhetischen Prinzipien orientiert. Afros und Schryer (2009) kommen zu einem ähnlichen Ergebnis bezüglich der Verwendung von „promotional metadiscourse“ und attestieren sogar verschwimmende Grenzen mit den literarischen Texten selbst (305). Die Studierenden von Studiengängen wie „Germanistik“ finden sich also mit (mindestens) zwei unterschiedlichen Fachkulturen und Sprachen konfrontiert, deren Erwerb überwiegend auf dem Weg der Imitation erfolgt (Graefen 1999). Dieser Beitrag hat das Ziel, die stilistischen Unterschiede zwischen den beiden Fächern mithilfe einer n-Gramm-Analyse zu beschreiben und damit bei Lehrenden und Studierenden zu einem höheren Bewusstsein für die damit verbundenen Herausforderungen beizutragen.

Methode

Die hier verwendete Methode ist eine n-Gramm-Analyse, die (fast) rein datengeleitet funktioniert und keine spezifischen Hypothesen erfordert. Ein n-Gramm ist eine Sequenz aus *n* Elementen, im einfachsten Fall aus Wörtern. N-Gramm-basierte Verfahren sind insbesondere in der Computerlinguistik verbreitet, wenn es um einfach zu berechnende Modellierungen von Sprache geht (Jurafsky und Martin 2009). Auch für die linguistische Interpretation wurden n-Gramme bereits genutzt: Scharloth und Bubenhofer (2012) zeigen bei der Analyse von Tonbandprotokollen zweier 68-Kommunen, dass auf diese Weise charakteristische Muster identifiziert werden können, die mit außersprachlichen Merkmalen der beiden Gruppen in Verbindung gebracht werden können. Mahlberg (2013) nutzt ein ähnliches Vorgehen zur Charakterisierung der Prosa Charles Dickens’, Biber et al. (2004) beschreiben unterschiedliche Formen der Wissenschaftssprache anhand sog. lexical bundles. Mit Ausnahme von Scharloth und Bubenhofer (2012) wird in diesen Ansätzen nur die Tokenebene einbezogen. Im Rahmen des hier präsentierten Vorhabens sollen die Potentiale zusätzlicher syntaktischer Informationen ermittelt werden.

Die Datengrundlage der folgenden Analyse ist ein Korpus aus 60 deutschen Dissertationen (30 pro Fach, ca. 3,5 Mio. Token) aus dem Zeitraum von 2003 bis 2016, die an 15 unterschiedlichen deutschen Universitäten eingereicht wurden und über universitäre Server online zur Verfügung stehen. Im Rahmen der Datenaufbereitung wurden semiautomatisch Textelemente ausgeschlossen, die nicht zur Zielvarietät gehören (Zitate), nicht aus Fließtext bestehen (Tabellen, Abbildungen,...) oder den Textfluss unterbrechen (Fußnoten). Die Texte wurden außerdem automatisch mit Informationen zu Lemma, Wortart und syntaktischen Dependenzstrukturen annotiert. [1]

Aus diesen Daten wurden n-Gramme der Größe n = 1 bis 5 generiert, die aus Token bzw. Wortartentags bestehen. Neben traditionellen, linearen n-Grammen, die der Reihenfolge der Wörter an der Textoberfläche folgen, wurden zusätzlich syntaktische n-Gramme generiert, die der Dependenzstruktur im Satz folgen (siehe Abbildung 1, beschrieben von Sidorov et al. 2012, Goldberg und Orwant 2013). Die Frequenzen aller n-Gramme mit mindestens zehn Vorkommen (rund 500.000) wurden signifikanzbasiert mit dem t-Test verglichen (siehe Empfehlungen in Lijffijt et al. 2014, Paquot und Bestgen 2009). Die Auswertung bezieht sich auf die n-Gramme mit den größten Unterschieden zwischen den linguistischen und literaturwissenschaftlichen Texten.

Abbildung 1: Lineare und syntaktische n-Gramme im Vergleich an einem Beispielsatz (vgl. Andresen und Zinsmeister 2017)

Ergebnisse

Exemplarisch werden hier die Ergebnisse zu den Wortarten-Unigrammen sowie den linearen und syntaktischen Token-Trigrammen präsentiert. Abbildung 2 zeigt die 20 Wortarten [2] mit den größten Unterschieden zwischen den beiden Disziplinen. Je weiter außen sich die Wortart befindet, desto größer ist der Unterschied. Wortarten auf der rechten Seite sind in der Literaturwissenschaft, die auf der rechten in der Linguistik häufiger. Der mit Abstand größte Unterschied zeigt sich in den attributiv gebrauchten Possessivpronomen (PPOSAT, z. B. *seine Existenz* (Lit_Stu_30 [3])). Verwandt hiermit

sind Unterschiede in den Reflexivpronomen (PRF) und Personalpronomen (PPER). Zusammen mit einer ebenfalls deutlich höheren Frequenz von Eigennamen (NE) spiegelt sich hier, dass sich literaturwissenschaftliche Texte in weitaus höherem Maße als die Linguistik mit Personen beschäftigen, seien es reale Autor/inn/en oder literarische Figuren, zum Beispiel:

(1) So bildet ihr autobiographisches Werk eine Brücke zwischen Tradition und Moderne. (Lit_Stu_30)

Weitere Unterschiede zeigen sich im Zusammenhang mit der Verbverwendung: Bei den finiten Verben der literaturwissenschaftlichen Texte handelt es sich eher um Vollverben (VVFIN), während finite Modal- und Auxiliarverben [4] (VMFIN, VAFIN) in der Linguistik frequenter sind. Korrespondierend dazu sind Auxiliarverben im Infinitiv (insb. *werden* in Passivkonstruktionen mit Modalverb) und Vollverben in ihrer Partizipform (VVPP) in der Linguistik häufiger. Lediglich die Form des passiven Perfekts ist in der Literaturwissenschaft häufiger, wie das Tag VAPP zeigt. Insgesamt lässt sich sagen, dass die Sprache der Linguistik im Vergleich mit der Literaturwissenschaft durch komplexe Verbkonstruktionen gekennzeichnet ist.

In der Linguistik zeigt sich außerdem eine höhere Frequenz von attribuierenden Indefinitpronomen (PIAT). Darunter sind die Lemmata *kein*, *aller* und *beide* am häufigsten. Dies kann damit in Verbindung gebracht werden, dass die Linguistik in stärkerem Maße auf Generalisierungen abzielt. Die höhere Frequenz von Zahlen (CARD) in der Linguistik überrascht nicht, da quantitative Verfahren hier deutlich häufiger zum Einsatz kommen als in der Literaturwissenschaft. Das Tag PRELS (Relativpronomen) weist auf eine häufigere Verwendung von Relativsätzen in der Literaturwissenschaft hin.

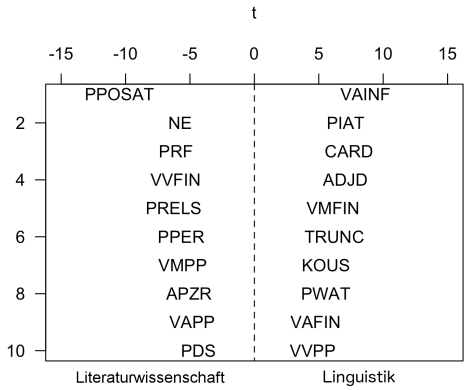

Abbildung 2: Die distinktivsten Wortarten, visualisiert anhand der Teststatistik t. Visualisierung inspiriert durch das R-Paket *stylo* (Eder et al. 2016).

Ergänzend werden in Tabelle 1 Informationen auf Wortebene herangezogen, zunächst ohne zusätzliche syntaktische Information. Gezeigt werden die zehn distinktivsten linearen Trigramme (inkl. Interpunktion). Hier spiegeln sich viele der Phänomene, die bereits auf Ebene der Wortarten erkennbar waren. Die für die Literaturwissenschaft charakteristischen Muster scheinen mehrheitlich aus Relativsätzen zu stammen. Dabei ist zu bedenken, dass der Anfang von Nebensätzen besonders leicht durch eine n-Gramm-Analyse erfasst werden kann, da hier nur ein begrenztes Maß an Variation möglich ist. Das klarste Trigramm für die Linguistik hingegen weist auf die bereits beschriebene häufigere Verwendung von Passiv und Modalverben hin, hier speziell in Kombination miteinander. Interessant ist das n-Gramm *die bei der*, das in 36 von 43 Fällen aus einem Relativsatz stammt, aber in der Linguistik häufiger ist. Mit dem Relativpronomen *die* kann es sich auf Feminina beziehen, in der Mehrzahl handelt es sich im Korpus aber um Substantive im Plural. Das passt zu der bereits oben genannten Annahme, dass die Literaturwissenschaft sich tendenziell exemplarisch mit konkreten Einzelphänomenen beschäftigt, die Linguistik hingegen in stärkerem Maße Generalisierungen anstrebt, die den Plural wahrscheinlich machen.

Rang	n-Gramm	Häufiger im Fach
1	, der sich	Literaturwissenschaft
2	, der seine	Literaturwissenschaft
3	, das sich	Literaturwissenschaft
4	werden können .	Linguistik
5	, die ihm	Literaturwissenschaft
6	, der die	Literaturwissenschaft
7	, dass eine	Linguistik
8	, vor dem	Literaturwissenschaft
9	, die er	Literaturwissenschaft
10	die bei der	Linguistik

Tabelle 1: Die distinktivsten linearen Token-Trigramme

Tabelle 2 zeigt die häufigsten syntaktischen Trigramme, die zusätzlich Informationen zur Dependenzstruktur im Satz nutzen. „>“ zeigt hier ein syntaktisches Dominanzverhältnis an. Die Relativsatzmuster sind hier nicht vorhanden, da ihre gute Erkennbarkeit in der Analyse vermutlich primär auf der linearen Abfolge von Interpunktion, Relativpronomen und folgendem Wort beruht. Das Muster *können>werden>*. ist höher gerankt als das Gegenstück in der linearen Analyse, da hier nicht nur unmittelbar aufeinanderfolgende Instanzen erfasst werden, sondern auch solche mit Distanzstellung:

(2) Einige Substantive können nicht eindeutig einer Geschlechtskategorie zugeordnet werden [...]. (Lin_Bam_01)

Zusätzlich tauchen Kombinationen von Passiv mit dem Modalverb *müssen* und *können* mit *sein* auf. Viele der für die Literaturwissenschaft charakteristischen Muster haben eine direkte lineare Entsprechung: So steht das syntaktische n-Gramm *für>Leben>das* für die lineare

Abfolge *für das Leben*. Allerdings umfasst das syntaktische n-Gramm zusätzlich Instanzen, in denen beispielsweise das Substantiv noch durch Attribute modifiziert wird, z. B. *für das eigene Leben* (Lit_Jen_19).

Rang	n-Gramm	Häufiger im Fach
1	können>werden>.	Linguistik
2	in>Regel>der	Linguistik
3	und>können>werden	Linguistik
4	in>Vorstellung>der	Literaturwissenschaft
5	müssen>werden>,	Linguistik
6	in>Darstellung>der	Literaturwissenschaft
7	für>Leben>das	Literaturwissenschaft
8	ist>Wunsch>der	Literaturwissenschaft
9	mit>Realität>der	Literaturwissenschaft
10	können>sein>.	Linguistik

Tabelle 2: Die distinktivsten syntaktischen Token-Trigramme

Fazit

In der hier präsentierten Analyse konnten deutliche stilistische Unterschiede zwischen Linguistik und Literaturwissenschaft gezeigt werden. Die Linguistik zeichnet sich demzufolge durch komplexe Verben (Passiv und Modalverben), die stärkere Verwendung von Zahlen sowie Mustern der Generalisierung aus. In der Literaturwissenschaft finden sich mehr Bezüge auf Personen und komplexe Nominalphrasen mit Relativsätzen.

Die verwendete Methode hat stark explorativen Charakter, sodass viele der hier angebotenen Interpretationen zunächst als Hypothesen betrachtet werden sollten und einer sorgfältigen Prüfung in Folgestudien bedürfen. Zusätzlich ergibt sich mit der Methode eine Beschränkung auf Phänomene, die sich auf konstante Weise auf der sprachlichen Oberfläche niederschlagen.

Im Rahmen dieses Beitrags wurden nur besonders hoch gerankten n-Gramme betrachtet und interpretiert. Andresen und Zinsmeister (2017) präsentieren ergänzend ein Annotationsexperiment, in dessen Rahmen insgesamt 420 Token- und Wortarten-n-Gramme auf die enthaltenen linguistischen Informationen hin ausgewertet wurden. In Folgearbeiten gilt es das Potential unterschiedlicher n-Gramm-Typen für eine Stilanalyse zu erforschen. Der Fokus wird dabei auf stärker syntaktisch informierten Formen liegen, die beispielsweise Informationen zu Token und Wortart kombinieren oder die syntaktischen Dependenzrelationen zwischen den Elementen einbeziehen.

Fußnoten

1. mit einem auf der Dependenzversion des TIGER-Korpus (Seeker und Kuhn 2012) trainierten Modell für MATE (Bohnet 2010)
2. Die Wortartentags stammen aus dem STTS (Schiller et al. 1999).
3. Die Bezeichnung der Korpustexte setzt sich aus einem Kürzel für die Disziplin, die Universität und einer fortlaufenden Zahl zusammen.
4. Bei den Verben *haben*, *sein* und *werden* wird nicht zwischen einer Verwendung als Auxiliar- oder Vollverb unterschieden.

Bibliographie

Afros, Elena / Schryer, Catherine F. (2009): „Promotional (meta)discourse in research articles in language and literary studies“, in: *English for Specific Purposes*. 28 (1), 58–68, doi: 10.1016/j.esp.2008.09.001.

Andresen, Melanie / Zinsmeister, Heike (2017): „Approximating Style by N-gram-based Annotation“, in: *Proceedings of the Workshop on Stylistic Variation*. Copenhagen, Denmark: Association for Computational Linguistics 105–115.

Biber, Douglas / Conrad, Susan / Cortes, Viviana (2004): „If you look at…: Lexical bundles in university teaching and textbooks“, in: *Applied linguistics*. 25 (3), 371–405.

Bohnet, Bernd (2010): „Very High Accuracy and Fast Dependency Parsing is not a Contradiction“, in: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): Stylometry with R: A Package for Computational Text Analysis. The R Journal 8(1). 107–121.

Goldberg, Yoav / Orwant, Jon (2013): „A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books“, in: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Atlanta, Georgia, USA. 241–247.

Graefen, Gabriele (1999): „Wie formuliert man wissenschaftlich?“, in: Barkowski, Hans; Wolff, Armin (Hrsg.) *Alternative Vermittlungsmethoden und Lernformen auf dem Prüfstand. Wissenschaftssprache - Fachsprache. Landeskunde aktuell. Interkulturelle Begegnungen - interkulturelles Lernen*. Regensburg: Fachverband Deutsch als Fremdsprache (Materialien Deutsch als Fremdsprache), 222–239.

Haggan, Madeline (2004): „Research paper titles in literature, linguistics and science: dimensions of attraction“, in: *Journal of Pragmatics*. 36 (2), 293–317, doi: 10.1016/S0378-2166(03)00090-0.

Jurafsky, Dan / Martin, James H. (2009): *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2. Aufl. Upper Saddle River, London: Pearson (Prentice Hall series in artificial intelligence).

Lijffijt, Jefrey / Nevalainen, Terttu / Säily, Tanja / Papapetrou, Panagiotis / Puolamäki, Kai / Mannila, Heikki (2014): „Significance testing of word frequencies in corpora“, in: *Digital Scholarship in the Humanities*. 1–24, doi: 10.1093/llc/fqu064.

Mahlberg, Michaela (2013): *Corpus stylistics and Dickens's fiction*. New York: Routledge (Routledge advances in corpus linguistics).

Paquot, Magali / Bestgen, Yves (2009): „Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction“, in: Jucker, Andreas H.; Schreier, Daniel; Hundt, Marianne (Hrsg.) *Corpora: Pragmatics and Discourse*. Brill 247–269, doi: 10.1163/9789042029101_014.

Scharloth, Joachim / Bubenhofer, Noah (2012): „Datengeleitete Korpuspragmatik. Korpusvergleich als Methode der Stilanalyse“, in: Felder, Ekkehard; Müller, Marcus; Vogel, Friedemann (Hrsg.) *Korpuspragmatik: thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin [u.a.]: De Gruyter (Linguistik - Impulse & Tendenzen), 195–230.

Schiller, Anne / Teufel, Simone / Thielen, Christine / Stöckert, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. Stuttgart, Tübingen. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [letzter Zugriff am 11.01.2018].

Seeker, Wolfgang / Kuhn, Jonas (2012): „Making Ellipses Explicit in Dependency Conversion for a German Treebank“, in: *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey 3132–3139.

Sidorov, Grigori / Velasquez, Francisco / Stamatatos, Efstathios / Gelbukh, Alexander / Chanona-Hernández, Liliana (2012): „Syntactic Dependency-Based N-grams as Classification Features“, in: Batyrshin, Ildar; Mendoza, Miguel González (Hrsg.) *Advances in Computational Intelligence*. Springer (Lecture Notes in Computer Science), 1–11, doi: 10.1007/978-3-642-37798-3_1.