Dependenzbasierte syntaktische Komplexitätsmaße

Proisl, Thomas
thomas.proisl@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg

Evert, Stefan
stefan.evert@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg

Die Beschreibung der Komplexität von (literarischen) Texten muss für jeden Aspekt, also Vokabular, Satz/Syntax, uneigentliche Rede, Intertextualität usw., gesondert vorgenommen werden. Im Folgenden beschäftigen wir uns mit dem Aspekt Satz/Syntax, der lange Zeit vor allem über die durchschnittliche Satzlänge erfasst wurde (Sherman 1893, Flesch 1948, Best 2005). Dabei bleibt aber, so eine naheliegende Vermutung, die interne syntaktische Komplexität eines Satzes unberücksichtigt. Die meisten Leser würden z. B. einen stark verschachtelten Satz als syntaktisch komplexer einstufen als eine gleich lange parataktische Konstruktion. Unsere Arbeit zielt darauf, diesen Aspekt unter Verwendung der im *Natural Language Processing* (NLP) weitverbreiteten dependenzbasierten Syntaxmodelle messen zu können. Kontext unserer Überlegungen ist das Unterfangen, Textkomplexität quantitativ zu erfassen. So können Annahmen in der Literaturwissenschaft und Linguistik über die unterschiedliche Komplexität der Texte bestimmter Gattungen, Autoren oder gar von Teilsystemen, z. B. populäre Literatur vs. Hochliteratur, empirisch überprüft werden. Bislang wird die syntaktische Komplexität überwiegend auf Phrasenstrukturbäumen ermittelt (für eine Übersicht siehe Vajjala Balakrishna 2015: 51–52), allerdings fehlen dafür in vielen Sprachen verlässliche NLP-Werkzeuge. Auf der anderen Seite stehen mit dem Universal-Dependencies-Projekt (Nivre u. a. 2016)[1] bereits mehr als 100 manuell erstellte Baumbanken in über 60 Sprachen (darunter auch ältere Sprachstufen) in einer sprachübergreifend konsistenten Annotation zur Verfügung und es gibt computerlinguistische Pipelines wie etwa UDPipe (Straka und Straková 2017),[2] die Texte in allen diesen Sprachen tokenisieren, taggen, lemmatisieren und parsen können. Von daher liegt es nahe, die syntaktische Komplexität von Texten auch mit dependenzbasierten Maßen zu messen. Einen ersten Vergleich von dependenzbasierten Komplexitätsmaßen hat Oya (2012) durchgeführt.

Für unsere Untersuchung verwenden wir ein deutschsprachiges Korpus von knapp 1.000 Romanen aus den letzten 60 Jahren. Bei etwa 85% der Texte handelt es sich um Heftromane (Romanzen (13%), Science Fiction (65%) und Horror (7%)), bei den restlichen 15% um Hochliteratur (kanonische Texte und/oder Literaturpreisträger). Alle Texte wurden mit dem DARIAH-DKPro-Wrapper (Jannidis u. a. 2016)[3] verarbeitet.

Syntaktische Komplexitätsmaße sind typischerweise auf Satzebene definiert. Wir berechnen für jeden Satz die folgenden Maße:[4]

- *Average dependency distance* (= durchschnittlicher Abstand zweier durch eine Dependenzrelation verbundener Tokens (Liu 2008; Oya 2011))
- *Closeness centrality* des Wurzelknotens (= Kehrwert der durchschnittlichen Länge der kürzesten Pfade vom Wurzelknoten zu allen anderen Knoten); hier bedeutet ein kleinerer Wert eine höhere Komplexität
- *Closeness centralization* (= Erweiterung der closeness centrality von einem einzelnen Knoten auf einen ganzen Graphen (Freeman 1978)); hier bedeutet ein kleinerer Wert eine höhere Komplexität
- *Outdegree centralization*, die Erweiterung der *outdegree centrality* (= Anzahl der von einem Knoten ausgehenden Kanten) von einem einzelnen Knoten auf einen ganzen Graph (Freeman 1978); hier bedeutet ein kleinerer Wert eine höhere Komplexität
- Durchschnittliche Anzahl von Dependenten pro Token
- Höhe des Dependenzbaums (= der längste kürzeste Pfad vom Wurzelknoten zu einem anderen Knoten)

Zum Vergleich ermitteln wir zusätzlich die Satzlänge, d. h. die Anzahl Tokens pro Satz. Um einen Wert für die syntaktische Komplexität eines gesamten Textes zu erhalten, bilden wir jeweils die Mittelwerte über alle Sätze.

Die Ergebnisse sind in den folgenden sechs Grafiken als Boxplots dargestellt (der weiße Kreis markiert zusätzlich das arithmetische Mittel):

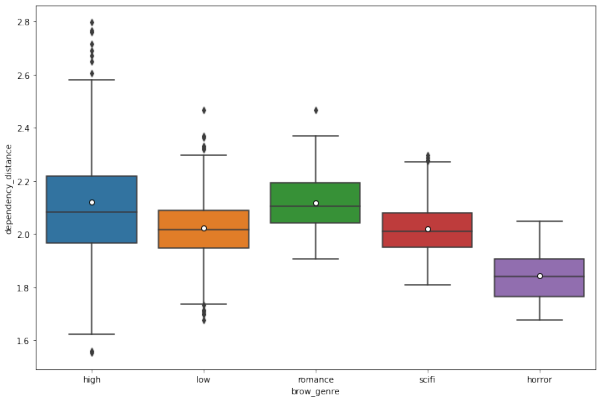

Abbildung 1. Average dependency distance

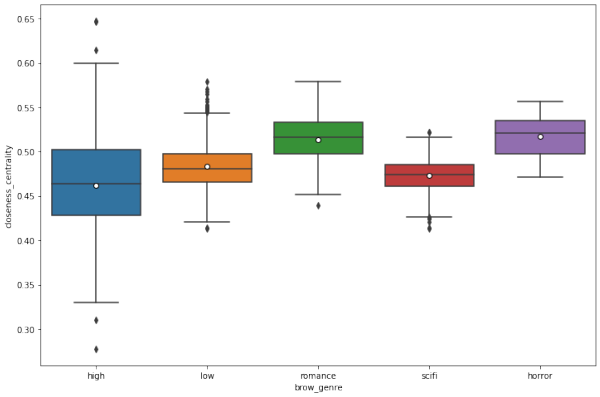

Abbildung 2. Closeness centrality

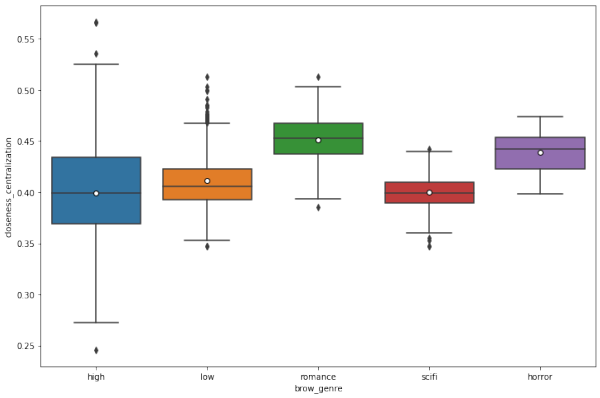

Abbildung 3.Closeness centralization

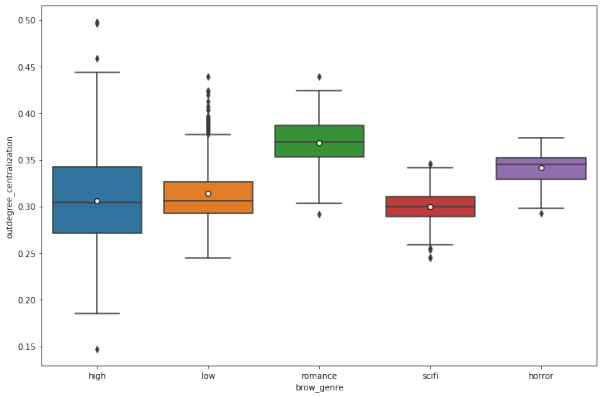

Abbildung 4. Outdegree centralization

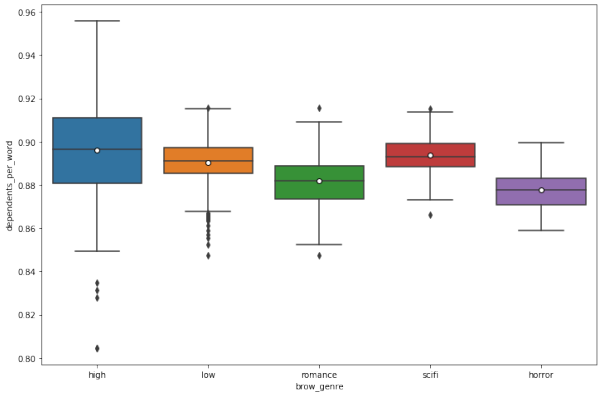

Abbildung 5. Dependenten pro Token

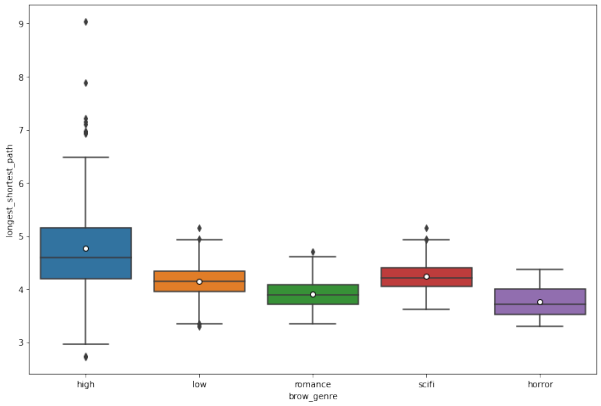

Abbildung 6. Höhe des Dependenzbaums

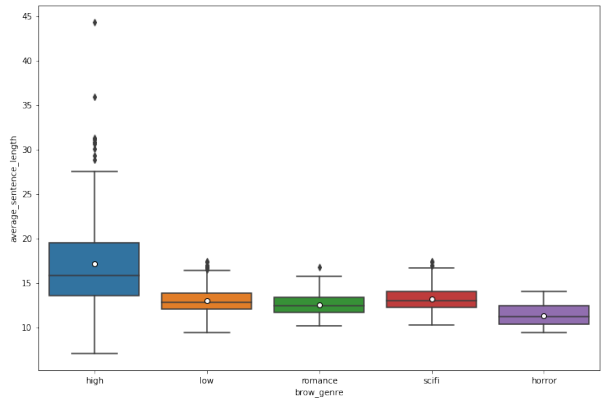

Abbildung 7. Satzlänge

Die Boxplots für Hoch- und Schemaliteratur insgesamt würden nahelegen, dass es für die untersuchten Maße keinen statistisch signifikanten Unterschied zwischen Hoch- und Schemaliteratur gibt. Die Detailansicht für die einzelnen Unterkategorien der Schemaliteratur bringt jedoch Interessantes zu Tage. Zwischen den einzelnen Kategorien untereinander gibt es deutlich ausgeprägtere Unterschiede als zwischen Hoch- und Schemaliteratur insgesamt. Besonders auffällig ist, dass fast alle Maße eine signifikant höhere Komplexität für Science-Fiction-Literatur anzeigen als für Romanzen oder Horrorhefte. Wahrscheinlich liegt das daran, dass das SF-Teilkorpus aus Romanen der Serie ‚Perry Rhodan' besteht, der auch von literaturwissenschaftlicher Seite eine Sonderrolle innerhalb der Heftromane zugeschrieben wird (Nast 2017). Ebenfalls auffällig ist, dass alle Maße eine viel größere Streuung für die Hochliteratur aufweisen als für die zahlenmäßig viel stärker vertretene Schemaliteratur. Dafür bieten sich zwei Erklärungsmodelle an: a) die Hochliteratur besteht eigentlich aus mehreren Gattungen, die sich wiederum deutlich voneinander unterscheiden; b) der Unterschied lässt sich auf die unterschiedlichen Eigenschaften der literarischen Teilfelder zurückführen – In der Hochliteratur dominiert der Wert Variation/Überraschung, in den populären Genres der Wert Erwartbarkeit, wahrscheinlich sogar erzwungen durch Lektoren.

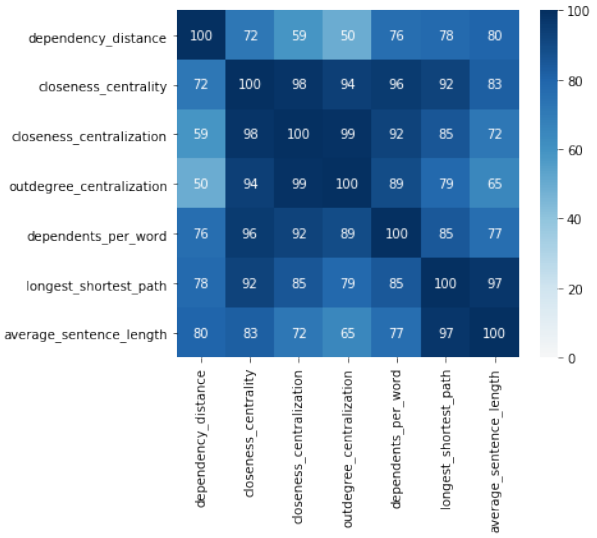

Abbildung 8.Pearson-Korrelationen zwischen den Komplexitätsmaßen

Eine Analyse der Pearson-Korrelationen zwischen den Komplexitätsmaßen zeigt, dass einige davon nahezu perfekt korrelieren.[5] So beispielsweise *closeness centralization* und *outdegree centralization* (r = 0,99), *closeness centrality* und *closeness centralization* (r = 0,98), die Höhe des Dependenzbaums und die Satzlänge (r = 0,97) und *closeness centrality* und die Anzahl Dependenten pro Wort (r = 0,96). *Average dependency distance* ist den anderen dependenzbasierten Maßen am unähnlichsten (0,50 # r # 0,78). Insgesamt betrachtet, korrelieren die verwendeten Maße zwar recht robust mit der durchschnittlichen Satzlänge (0,65 # r # 0,83), scheinen sich aber (mit Ausnahme der Höhe des Dependenzbaums) doch auch ausreichend stark von ihr zu unterscheiden um den durch das Parsen der Texte und Berechnen der dependenzbasierten Maße entstehenden Mehraufwand zu rechtfertigen. Zusätzlich könnte es durch die gezielte Entwicklung längenkorrigierter Maße gelingen, unterschiedliche Aspekte syntaktischer Komplexität getrennt voneinander zu erfassen.

Fußnoten

1. http://universaldependencies.org/
2. http://ufal.mff.cuni.cz/udpipe
3. https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper
4. https://github.com/tsproisl/Linguistic_and_Stylistic_Complexity
5. *Closeness centrality*, *closeness centralization* und *outdegree centralization* wurden für diese Analyse mit #1 multipliziert, damit für alle Maße größere Werte eine höhere Komplexität anzeigen.

Bibliographie

Best, Karl-Heinz (2005): *Satzlänge*, in: **Köhler, Reinhard / Altmann, Gabriel / Piotrowski, Rajmund G.:** *Quantitative Linguistik / Quantitative Linguistics*. Berlin: de Gruyter-Mouton 298–304.

Flesch, Rudolf (1948): *A New Readability Yardstick*, in: Journal of Applied Psychology 32 Nr. 3: 221–233. 10.1037/h0057532 [letzter Zugriff am 8. Januar 2019].

Freeman, Linton C. (1978): *Centrality in Social Networks. Conceptual Clarification*, in: Social Networks 1, Nr. 3: 215–239. 10.1016/0378-8733(78)90021-7 [letzter Zugriff am 15. Oktober 2018].

Jannidis, Fotis / Pernes, Stefan / Pielström, Steffen / Reger, Isabella / Reimers, Nils / Vitt, Thorsten (2016): *DARIAH-DKPro-Wrapper Output Format (DOF) Specification*, in: DARIAH-DE Working Papers 20. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2016-6-2 [letzter Zugriff am 15. Oktober 2018].

Liu, Haitao (2008): *Dependency Distance as a Metric of Language Comprehension Difficulty*, in: Journal of Cognitive Science 9, Nr. 2: 159–191. http://cogsci.snu.ac.kr/jcs/issue/vol9/no2/JCS_Vol_09_+No_2+p.+159+- +191+Dependency+Distance+as+a+Metric+of+Language+Comprehension+Difficulty.pdf [letzter Zugriff am 15. Oktober 2018].

Nast, Mirjam (2017): *„Perry Rhodan“ lesen. Zur Serialität der Lektürepraktiken einer Heftromanserie*. Bielefeld: transcript.

Nivre, Joakim / >de Marneffe, Marie-Catherine / Ginter, Filip / Goldberg, Yoav / Hajic, Jan / Manning, Christopher D. / McDonald, Ryan / Petrov, Slav / Pyysalo, Sampo / Silveira, Natalia / Tsarfaty, Reut / Zeman, Daniel (2016): *Universal Dependencies v1: A Multilingual Treebank Collection*, in: **Calzolari, Nicoletta / Choukri, Khalid / Declerck, Thierry / Goggi, Sara / Grobelnik, Marko / Maegaard, Bente / Mariani, Joseph / Mazo, Hélène / Moreno, Asunción / Odijk, Jan / Piperidis, Stelios (eds.):** *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association 1659–1666. http://www.lrec-conf.org/proceedings/lrec2016/pdf/348_Paper.pdf [letzter Zugriff am 15. Oktober 2018].

Oya, Masanori (2011): *Syntactic Dependency Distance as Sentence Complexity Measure*, in: Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics. 313–316. https://www.researchgate.net/profile/Masanori_Oya2/publication/266584664_Syntactic_Dependency_Distance_as_Sentence_Complexity_Measure/links/54fe480f0cf2672e223ed842.pdf [letzter Zugriff am 15. Oktober 2018].

Sherman, Lucius Adelno (1893): *Analytics of literature, a manual for the objective study of English prose and poetry*. Boston: Ginn. https://archive.org/details/analyticsofliter00sheruoft/page/n3 [letzter Zugriff am 8. Januar 2019].

Straka, Milan / Straková, Jana (2017): *Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe*, in: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver: Association for Computational Linguistics 88–99. http://www.aclweb.org/anthology/K17-3009 [letzter Zugriff am 15. Oktober 2018].

Vajjala Balakrishna, Sowmya (2015): *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Dissertation, Eberhard-Karls-Universität Tübingen. https://publikationen.uni-tuebingen.de/xmlui/bitstream/handle/10900/64359/THESIS-FINAL.pdf [letzter Zugriff am 15. Oktober 2018].