FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Dependenzbasierte syntaktische Komplexitätsmaße

Thomas Proisl* · Leonard Konle† · Stefan Evert* · Fotis Jannidis†

*FAU Erlangen-Nürnberg · †JMU Würzburg

Einleitung

- Verschiedene Aspekte der Komplexität (literarischer) Texte:
  - Vokabular
  - **Satz/Syntax**
  - Uneigentliche Rede
  - Intertextualität
  - . . .
- Syntaktische Komplexität:
  - Approximation durch oberflächennahe Merkmale, bspw. Satzlänge
  - Phrasenstrukturbäume
  - **Dependenzbäume**

Experimente

Dependenzbasierte Komplexitätsmaße

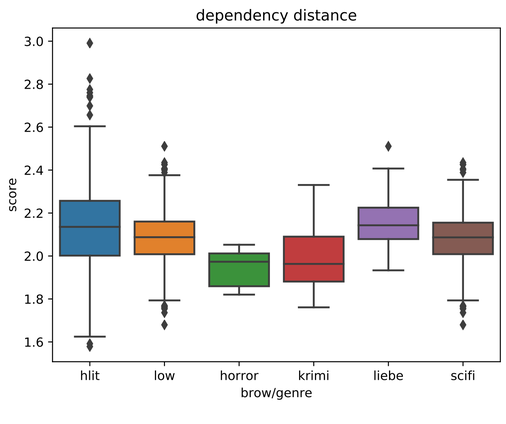
- Average dependency distance (= durchschnittlicher Abstand zweier durch Dependenzrelation verbundener Tokens)
- Closeness centrality des Wurzelknotens (= Kehrwert der durchschnittlichen Länge der kürzesten Pfade vom Wurzelknoten zu allen anderen Knoten)*
- Closeness centralization (= Erweiterung der closeness centrality auf ganzen Graph)*
- Outdegree centralization, Erweiterung der outdegree centrality (= Anzahl der von einem Knoten ausgehenden Kanten) auf ganzen Graph*
- Durchschnittliche Anzahl von Dependenten pro Token
- Höhe des Dependenzbaums (= längster kürzester Pfad vom Wurzelknoten zu einem anderen Knoten)
- Syntaktische Komplexität des Textes: **Mittelwert über Sätze**

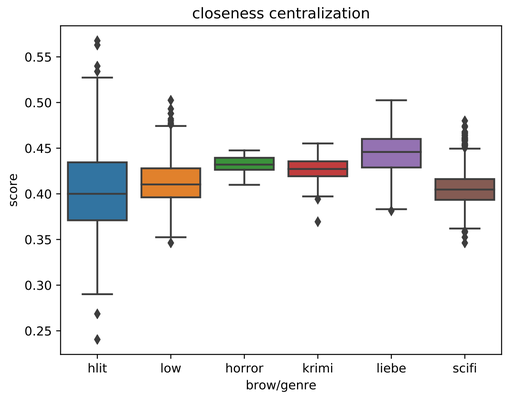
*Kleinerer Wert → höhere Komplexität

Korpus

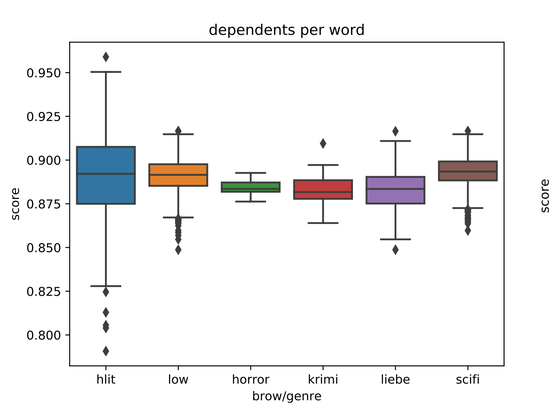
- Knapp 1.000 deutschsprachige Romane aus den letzten 60 Jahren
- Ca. 85% Heftromane (Romanzen (13%), Science Fiction (65%) und Horror (7%))
- Ca. 15% Hochliteratur (kanonische Texte und/oder Literaturpreisträger)
- Vorverarbeitung mit Kallimachos Preprocessing-Pipeline (https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/KallimachosEngines)

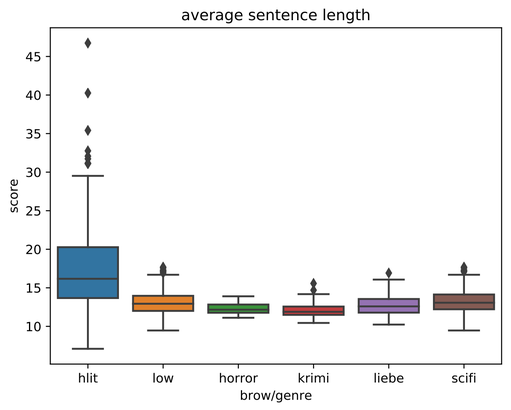
Ergebnisse



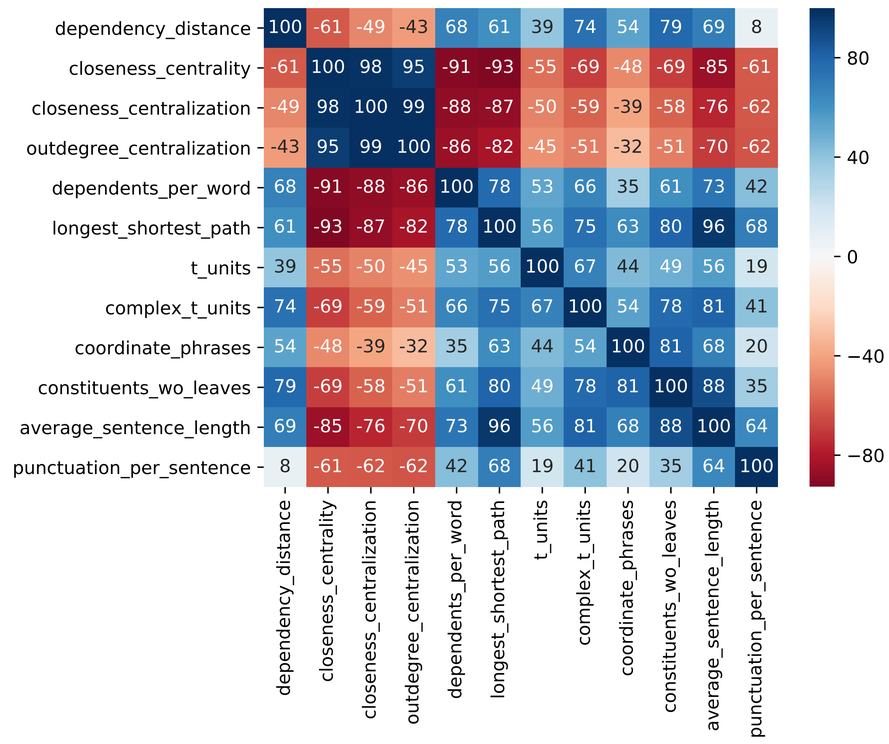


Ergebnisse





- Teils deutliche Gattungsunterschiede bei Heftromanen
- Höhere Komplexität von Science-Fiction → Sonderrolle der Perry-Rhodan-Reihe innerhalb der Heftromane (Nast 2017)
- Große Streuung bei Hochliteratur; mögliche Erklärungen:
  - Mehrere Gattungen mit deutlichen Unterschieden
  - Unterschiedliche Eigenschaften der literarischen Teilfelder: Variation/Überraschung (Hochliteratur) vs. Erwartbarkeit (Heftromane)



Spearman Rangkorrelationen zwischen den Maßen

- Satzlänge reflektiert Aspekte syntaktischer Komplexität (robuste Korrelationen mit allen Maßen)
- Dependenzbasierte Maße unterscheiden sich trotzdem deutlich von Satzlänge und konstituenzbasierten Maßen

Fazit und Ausblick

- Einzelnes Maß nicht ausreichend für zuverlässige Trennung von Hoch- und Schemaliteratur
- Trennung unterschiedlicher Aspekte syntaktischer Komplexität durch gezielte Entwicklung längenkorrigierter Maße?
- Classifier für Hoch- und Schemaliteratur?
  - Majority baseline: 85% accuracy
  - Kombination weniger Komplexitätsmaße (lexikalisch, dependenz- und konstituenzbasiert) erzielt 97% (SVM mit RBF-Kernel)

https://github.com/tsproisl/Linguistic_and_Stylistic_Complexity

{thomas.proisl, stefan.evert}@fau.de, {leonard.konle, fotis.jannidis}@uni-wuerzburg.de