Into the Perryverse

A CL Journey to the Realm of Lexical Complexity

Andreas Blombach, Thomas Proisl, Philipp Heinrich, Stefan Evert, Natalie Dykes

Chair of Computational Corpus Linguistics
Friedrich-Alexander-Universität Erlangen-Nürnberg

August 18–21, 2021

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Overview

1. Background: Kallimachos
2. Perry Rhodan
3. Complexity and the Vocabulary of Perry Rhodan
4. Complexity of Science Fiction
5. Conclusion and Outlook

Background: Kallimachos

Background: Kallimachos

BMBF project Würzburg/Erlangen

- Aim: Quantify the **surface complexity** of texts
  - `https://github.com/tsproisl/textcomplexity`
  - Main focus on richness or diversity of vocabulary
    → type-token ratio, Sichel's S, Yule's K, ...
- What is complexity?
  - Complexity $\neq$ readability
  - Analysis of strings, not psychological experiments
  - Structural property of a text, not mental processes

> *Complexity is first and foremost a matter of the number and variety of an item's constituent elements and of the elaborateness of the interrelational structure, be it organizational or operational.*
>
> (Rescher, 1998: 1)

Complexity measures

Surface-based measures

- Length-based:
  - ☞ Simplest surface-based measures
  - Word length in characters
  - Sentence length in words or characters
- Variability:
  - ☞ Lexical diversity, variability of words used in text
  - Type-token ratio
  - Honoré’s H (Honoré, 1979)
  - …and many more, such as MTLD (McCarthy, 2005)
- Dispersion:
  - ☞ Burstiness of word distributions
  - Gries’ DP_{norm} (Gries, 2008; Lijffijt and Gries, 2012)
  - Kullback-Leibler divergence

Complexity measures

Measures based on linguistic analysis

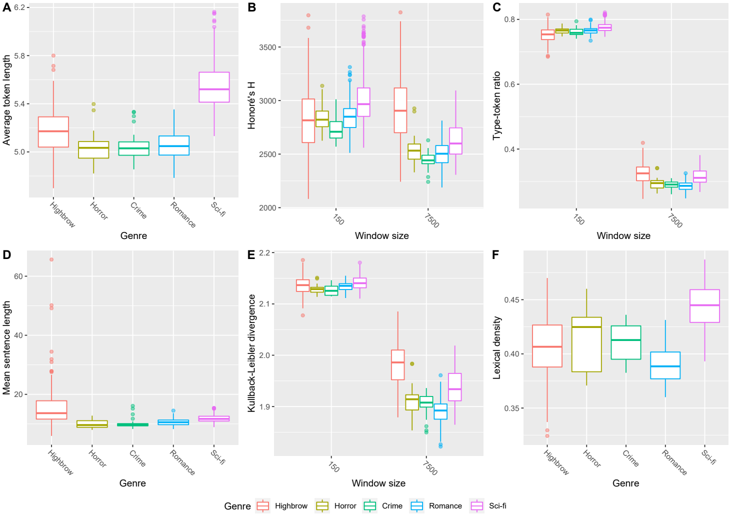
- Lexical density
  - ☞ Proportion of content words (based on POS tags)
- Rarity
  - ☞ Proportion of "rare" words (by comparison to the frequency list of some reference corpus)
- Semantic disparity
  - ☞ Repetition not only of identical, but of similar types: "degree of differentiation between lexical types in a text" (Jarvis, 2013: 25)
- Dependency-based:
  - ☞ Complexity of syntactic structures
  - ▸ Dependency distance (Oya, 2011)
  - ▸ Dependents per word

Stability of measures

- Almost all measures of variability and dispersion **dependent** on **text length**
- Solution: compute measures on windows of fixed size; score for text = arithmetic mean
- Caveat: chosen **window size** can influence ranking of texts (most noticeable for highbrow literature) → micro- and macro-diversity?

The Unexpected Complexity of Sci-fi Dime Novels

Selected complexity measures by genre and window size (where applicable)



Where does this apparent complexity come from?

- In our corpus: sci-fi dime novels = Perry Rhodan
- *Is it just Perry Rhodan, or is it science fiction in general?*

A Closer Look at Perry Rhodan

The Perryverse

- **Perry Rhodan:** German science fiction series named after its hero
- Perry Rhodan *Heftromane* (dime/pulp novels)
  - Weekly booklets of ca. 60 pages since September 1961
  - More than 3100 *Heftromane* → the world's "biggest science fiction series"
  - Arc storyline structure: One arc (*Zyklus*) $\approx$ 50–100 *Hefte*
- Lots of additional products:
  - Spinoff series (*Atlan*, focusing on one of the main characters)
  - Mini series
  - Paperbacks (e. g. *Planetenromane*)
  - Hardcover editions of *Heftromane* (e. g. *Silberbände*)
  - Comics, audio books, movies
  - ...
- Perry Rhodan Neo *Heftromane* (reboot of the story, published biweekly in parallel to the main series)

The Perry Rhodan Library

https://www.reddit.com/r/books/comments/iq31r1/i_22_finally_collected_and_stored_the_whole_perry/

A Few Cover Illustrations

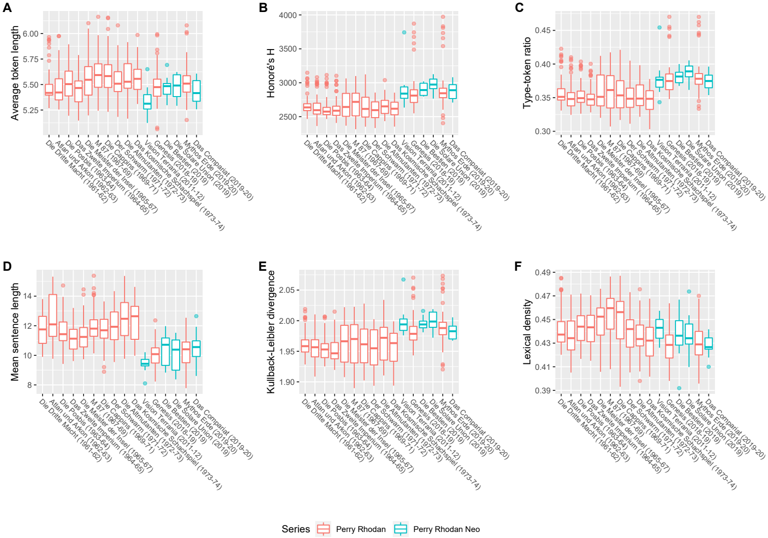
Complexity and the Vocabulary of Perry Rhodan

Corpus: Perry Rhodan

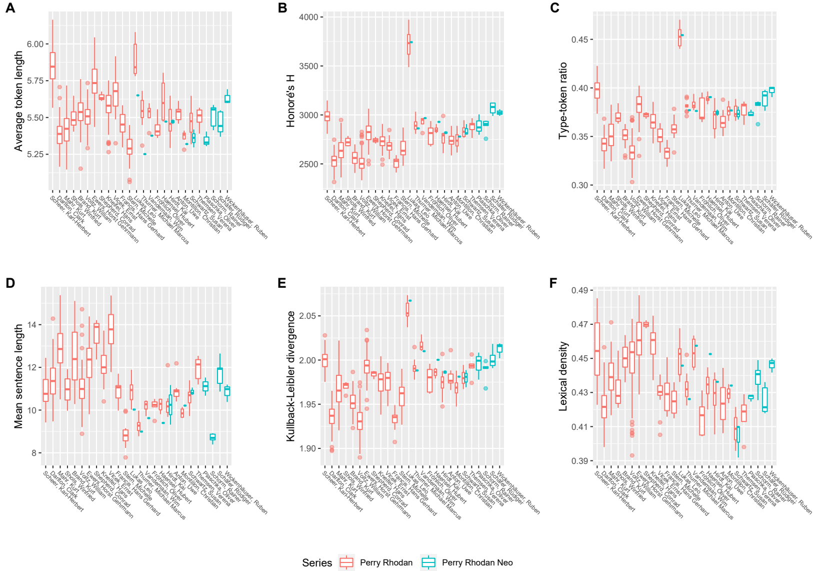
- 649 *Heftromane* from 1961 to 1974
- 100 *Heftromane* from 2018 to 2020 (kindly provided by the publisher)
- 37 longer *Heftromane* from the *Neo* series, from 2011 to 2020 (29 provided by the publisher)
- Expectation: newer novels, especially from the *Neo* series, less complex than older ones

Complexity over Time/*Zyklus*

Only a single window size from now on: 5,000 words

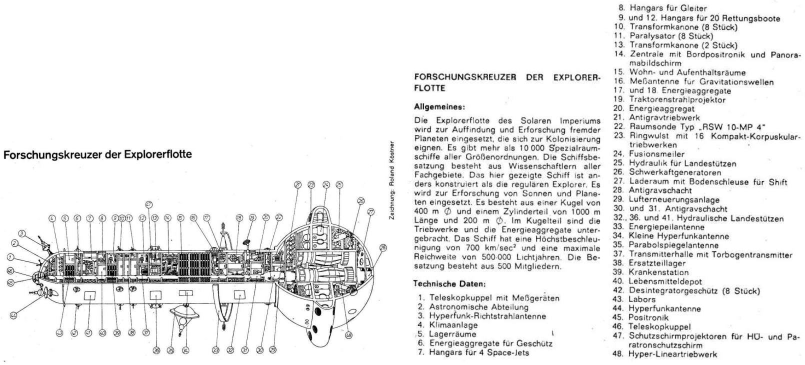


Complexity by Author



The Vocabulary of Perry Rhodan

Technical aspects are certainly important . . .



Forschungskreuzer der Explorerflotte

Zeichnung: Roland Köstner

FORSCHUNGSKREUZER DER EXPLORER-FLOTTE

Allgemeines:

Die Explorerflotte des Solaren Imperiums wird zur Auffindung und Erforschung fremder Planeten eingesetzt, die sich zur Kolonisierung eignen. Es gibt mehr als 10 000 Spezialraumschiffe aller Größenordnungen. Die Schiffsbesatzung besteht aus Wissenschaftlern aller Fachgebiete. Das hier gezeigte Schiff ist anders konstruiert als die regulären Explorer. Es wird zur Erforschung von Sonnen und Planeten eingesetzt. Es besteht aus einer Kugel von 400 m Ø und einem Zylinderteil von 1000 m Länge und 200 m Ø. Im Kugelteil sind die Triebwerke und die Energieaggregate untergebracht. Das Schiff hat eine Höchstbeschleunigung von 700 km/sec^2 und eine maximale Reichweite von 500.000 Lichtjahren. Die Besatzung besteht aus 500 Mitgliedern.

Technische Daten:

1. Teleskopkuppel mit Meßgeräten
2. Astronomische Abteilung
3. Hyperfunk-Richtstrahlantenne
4. Klimaanlage
5. Lagerräume
6. Energieaggregate für Geschütz
7. Hangars für 4 Space-Jets
8. Hangars für Gleiter
9. und 12. Hangars für 20 Rettungsboote
10. Transformkanone (8 Stück)
11. Paralysator (8 Stück)
13. Transformkanone (2 Stück)
14. Zentrale mit Bordpositronik und Panoramabildschirm
15. Wohn- und Aufenthaltsräume
16. Meßantenne für Gravitationswellen
17. und 18. Energieaggregate
19. Traktorenstrahlprojektor
20. Energieaggregat
21. Antigravtriebwerk
22. Raumsonde Typ „RSW 10-MP 4"
23. Ringwulst mit 16 Kompakt-Korpuskulartriebwerken
24. Fusionsmeiler
25. Hydraulik für Landestützen
26. Schwerkaftgeneratoren
27. Laderaum mit Bodenschleuse für Shift
28. Antigravschacht
29. Lufterneuerungsanlage
30. und 31. Antigravschacht
32., 36. und 41. Hydraulische Landestützen
33. Energiepeilantenne
34. Kleine Hyperfunkantenne
35. Parabolspiegelantenne
37. Transmitterhalle mit Torbogentransmitter
38. Ersatzteillager
39. Krankenstation
40. Lebensmitteldepot
42. Desintegratorgeschütz (8 Stück)
43. Labors
44. Hyperfunkantenne
45. Positronik
46. Teleskopkuppel
47. Schutzschirmprojektoren für HÜ- und Paratronschutzschirm
48. Hyper-Lineartriebwerk

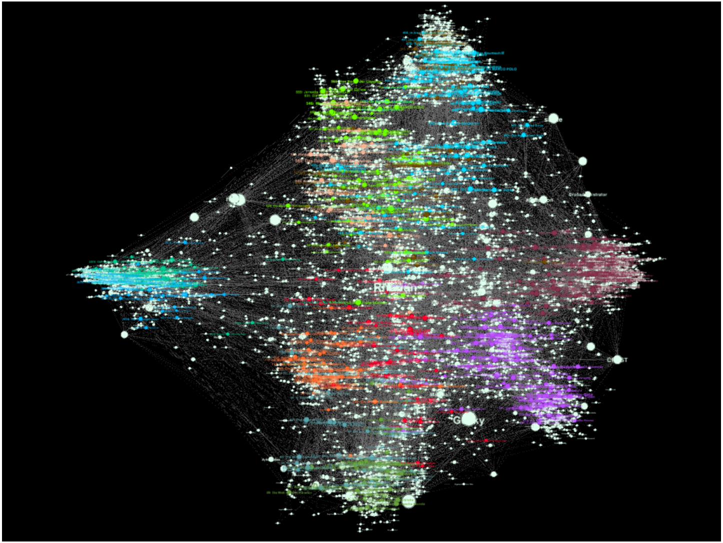
- Cutaway drawings (*Risszeichnungen*) like this are a staple of the series.
- Image source:
  `https://www.pr-materiequelle.de/riss/risszeichnung/r13_4.htm`

The Vocabulary of Perry Rhodan

- Perry Rhodan exhibits a great amount of idiosyncratic vocabulary, e. g. *Raumer*, *Impulsstrahler*, *Mausbiber*, *Arkonide*, *Linearraum*, *Zellaktivator*
- Use keyword analyses to extract vocabulary that
  1. characterizes Perry Rhodan as a whole (Perry Rhodan vs. reference corpus)
  2. characterizes individual *Heftromane* (one *Heft* vs. the rest)
- Can *Heft*-wise keywords be used to reconstruct storyline arcs (*Zyklen*)?

Keyword Network

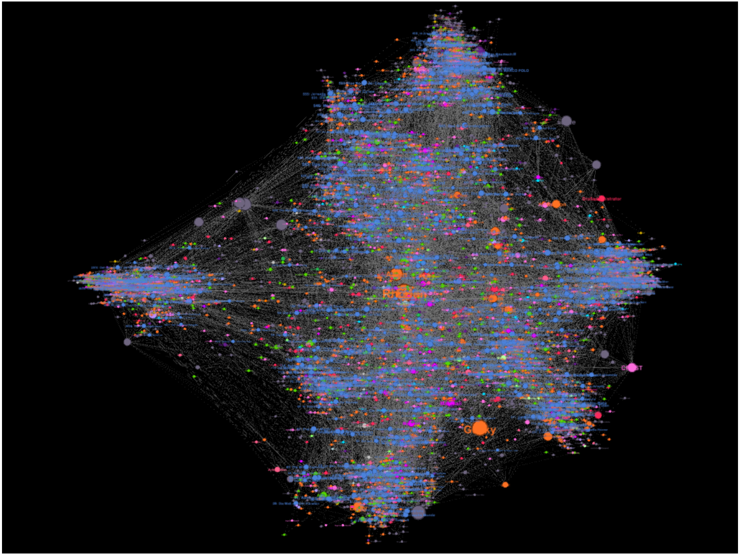
Nodes: individual issues and top 50 keywords for each; edge weights: LLR



Characterizing and Categorizing the Vocabulary (WIP)

- The *Perrypedia* is a wiki for the *Perryverse*, containing over 50,000 articles
- Scraped named entities and their categories using Scrapy
- Opens up new methods for analysis (distant reading)

Characterizing and Categorizing the Vocabulary (WIP)

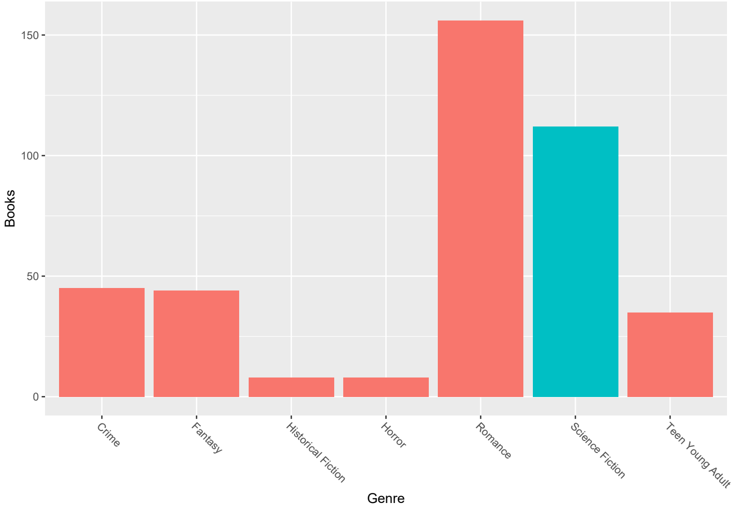


Complexity of Science Fiction

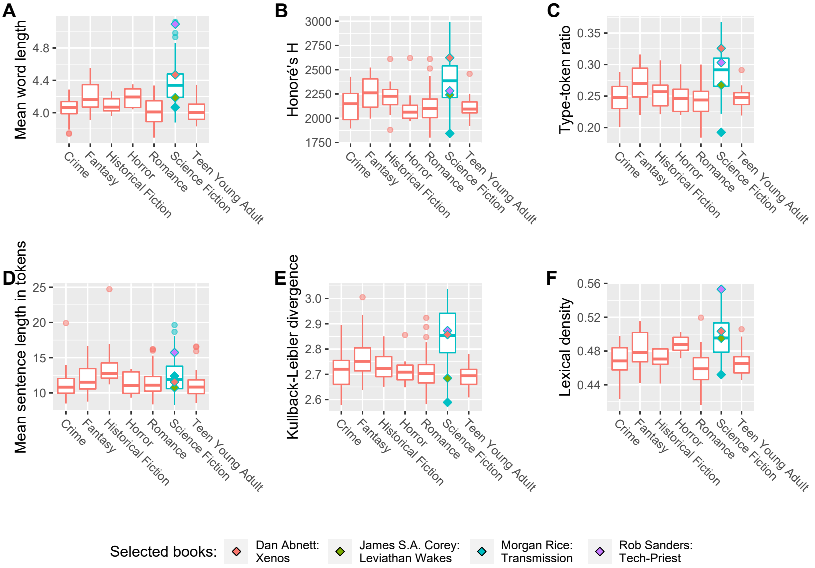
Is it just Perry Rhodan?

Our English E-book Corpus (29 million tokens)

Opportunistic collection via `https://www.bookbub.com` plus e-book bundles ⇒ currently very unbalanced

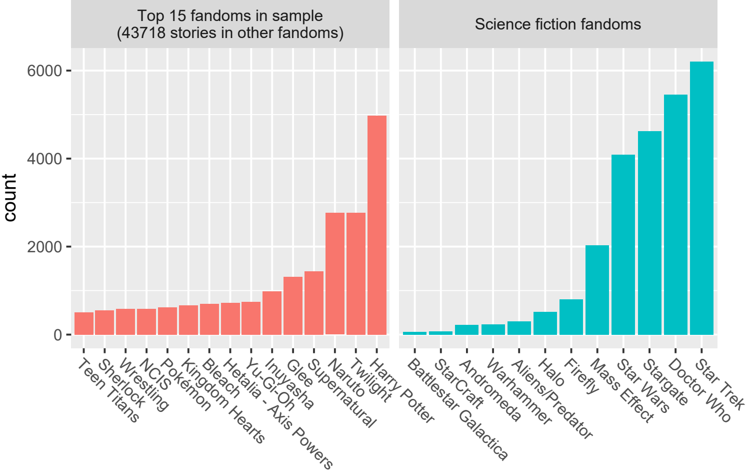


Complexity in an English E-book Corpus



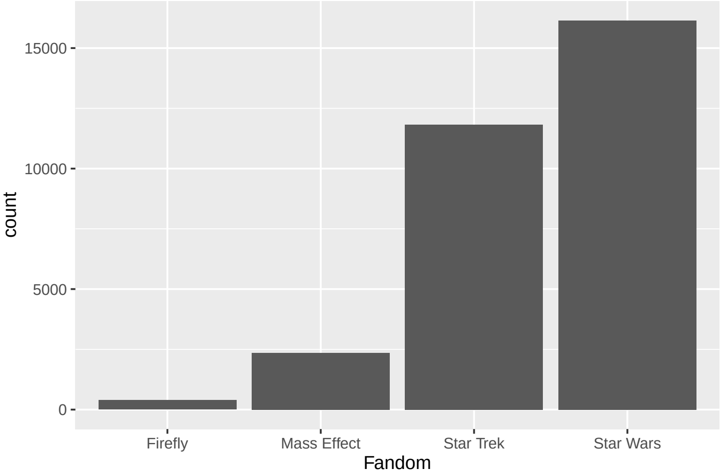
Corpus: fanfiction.net (2.3 billion tokens)

`https://archive.org/details/fanfictiondotnet_repack` (total corpus size >50 billion); science fiction fandoms are included in full, random sample for other fandoms



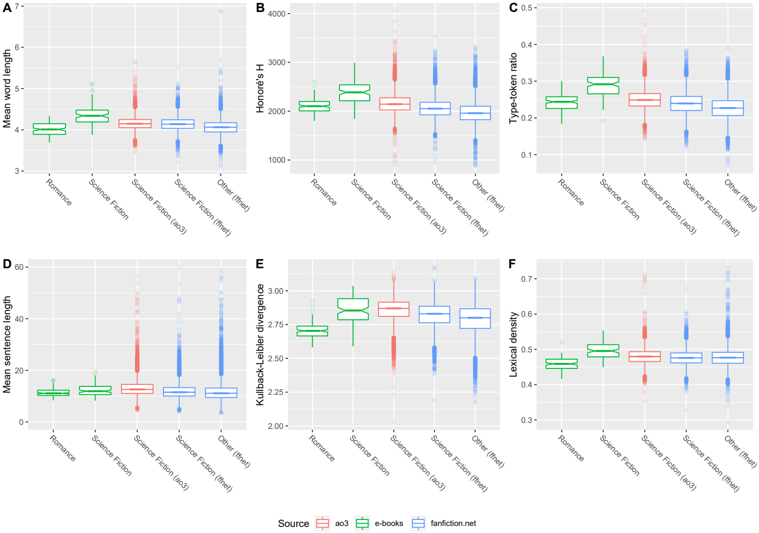
Corpus: AO3 (“Archive of Our Own”, 688 million tokens)

Collected using `https://github.com/radiolarian/AO3Scraper`



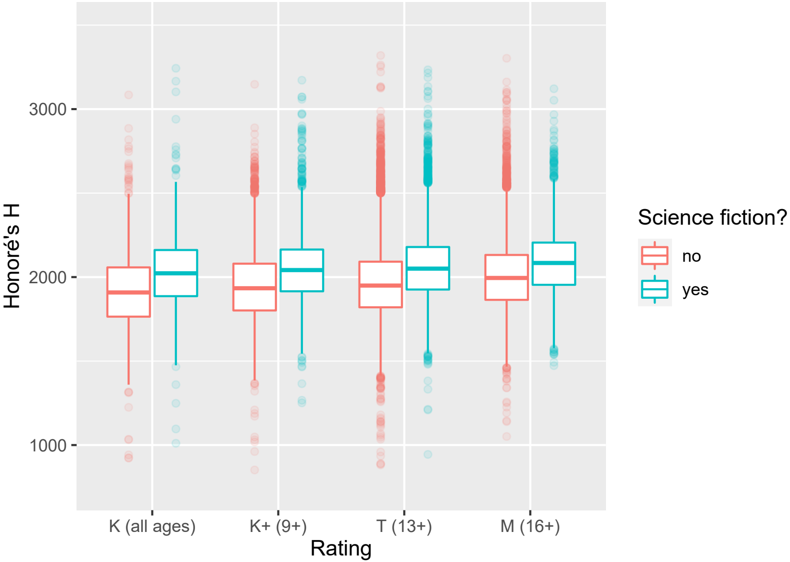
Complexity of Sci-fi Fanfiction

Is it just published sci-fi or also fan-written material?



But wait! Could writers and readers of science fiction fanfiction just be older on average?

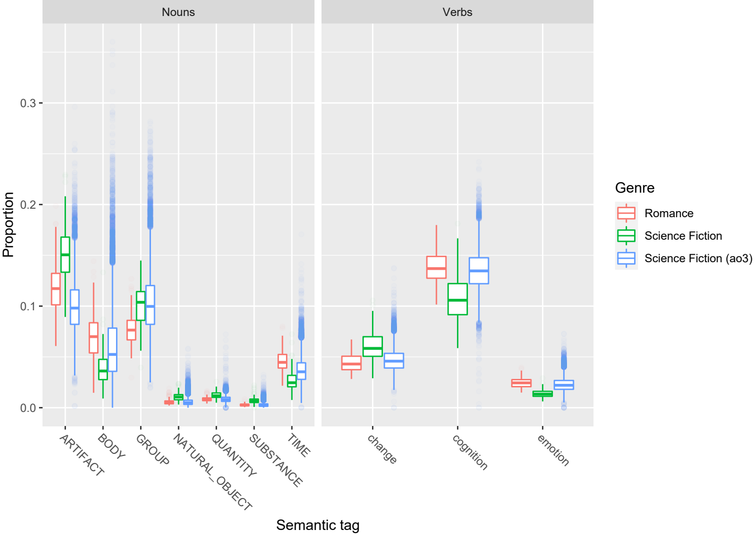
Complexity by rating (fanfiction.net)



But wait! Is sci-fi fanfiction really sci-fi?

Semantic tags: Romance vs. Sci-fi vs. Sci-fi Fanfiction

WordNet supersenses (25 noun, 15 verb); AMALGrAM 2.0



Conclusion and Outlook

Conclusions and Outlook

- It is not just Perry Rhodan
  - (English) published Sci-fi also greater vocabulary richness than other genres
  - Sci-fi fanfiction shows the same trend
  - Sci-fi fanfiction is partly "romance in disguise"
- Keywords able to capture important concepts and names
  - Successful reconstruction of *Zyklen*
  - *Perrypedia* as resource for categorizing the vocabulary

  ⇒ Distant reading informed by curated knowledge source
- Semantic tags surprisingly useful

References I

Stefan Th. Gries. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4):403–437, 2008.

Anthony Honoré. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177, 1979.

Scott Jarvis. Chapter 1. Defining and measuring lexical diversity. *Studies in Bilingualism*, page 13–44, 2013.

Jefrey Lijffijt and Stefan Th. Gries. Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics*, 17(1):147–149, 2012.

Phillip McCarthy. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD thesis, University of Memphis, 2005.

Masanori Oya. Syntactic dependency distance as sentence complexity measure. In *Proceedings of The 16th Conference of Pan-Pacific Association of Applied Linguistics*, pages 313–316, 2011.

Nicholas Rescher. *Complexity. A Philosophical Overview*. Routledge, 1998.

Thanks for listening.

Questions!