

# A Matter of Perspective

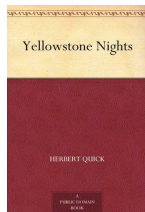
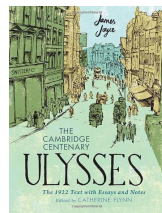
## Building a Multi-Perspective Annotated Dataset for the Study of Literary Quality

Yuri Bizzoni, Pascale Feldkamp, Mads Rosendahl Thomsen, Kristoffer L. Nielbo

Center for Humanities Computing, Aarhus University &  
Comparative Literature – Institute for Communication and Culture, Aarhus University

# Introduction

## What is “literary quality”?



### Text-extrinsic features:

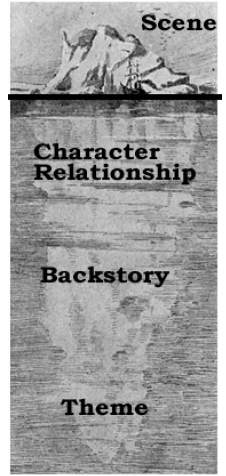
- Reader and critic demographics (Lassen et al., 2022; Koolen, 2018)
- An effect of reception and cultural dynamics (Bourdieu, 1993; Casanova, 2007)
- The cultural capital of ruling classes (Guillory, 1995)

### Text-intrinsic features:

- Authorial strategies and defamiliarization and foregrounding (Shklovsky, 1917; Mukarovsky, 1964)
- Narrative coherence and style (Bizzoni et al., 2023c,a)

# Related Works

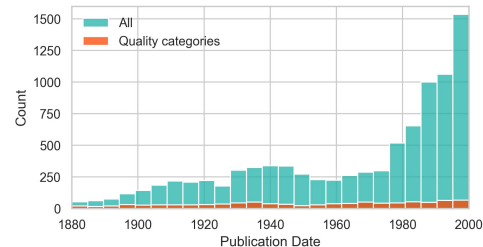
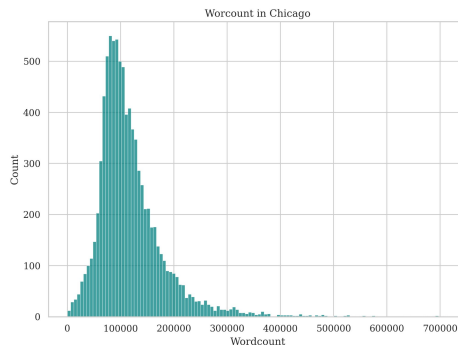
- Simplicity (Sherman 1893, Martin 1996, Garthwaite 2014)
- Economy - omitting needless words (King), or understating emotions (Hemingway)
- Affective aspects and narratology (Rebora 2023)
- Sentiment analysis from induced lexica (Islam et al. 2020)
- Changes in reader preferences as a historical development



# Corpus

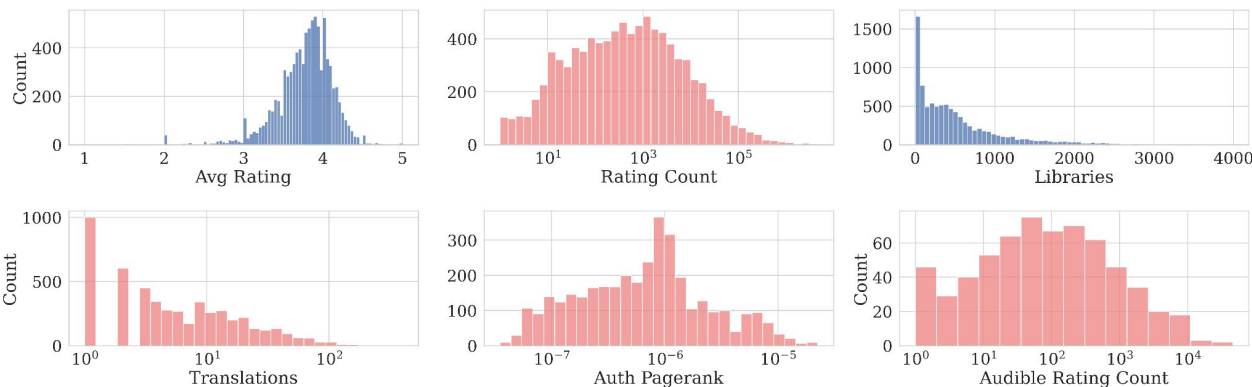
- Collected by Hoyt Long and Richard Jean So, using the number of library holdings for their selections
- High brow (Don DeLillo, Philip Roth) to genre fiction (Tolkien, Philip K. Dick)
- Anglophone authors with few exceptions
- We looked for titles associated with different quality proxies in our corpus

Titles	Authors	Titles per author
9088	3166	2.88



# Quality Metrics

- Most complicated from a conceptual standpoint
- Crowd-based (many unfiltered readers) and expert-based (often institutionally affiliated)
- Continuous and discrete, author-based and title-based



# Quality Metrics

## Crowd-Based Metrics:

- GoodReads, Audible (and StoryGraph)
- Wikipedia Author rank
- Best-selling titles

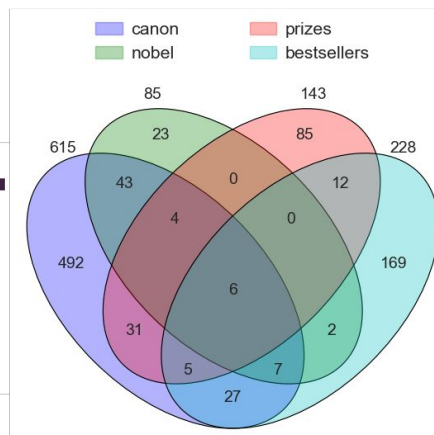
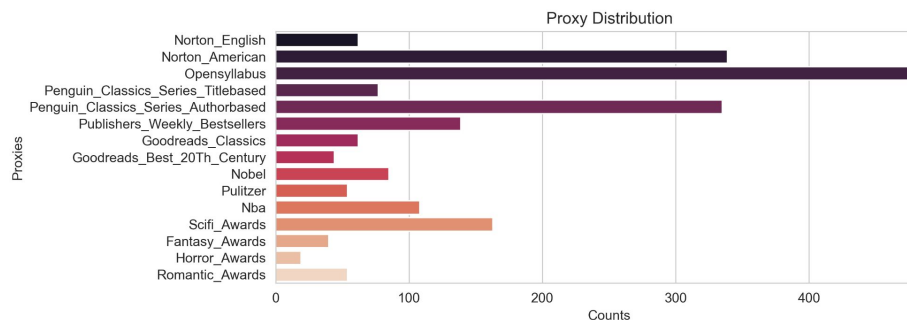


	Count
Libraries	9088
Translations	5082
PageRank*	3558
Audible Rat.Avg.	629
Audible Rat.Count	629
GR Rat.Avg.	8989
GR Rat.Count	8989

# Quality Metrics

## Expert-based Metrics:

- Literary Awards (more than 20, genre-specific and not)
- Canonical Anthologies
- University Syllabi

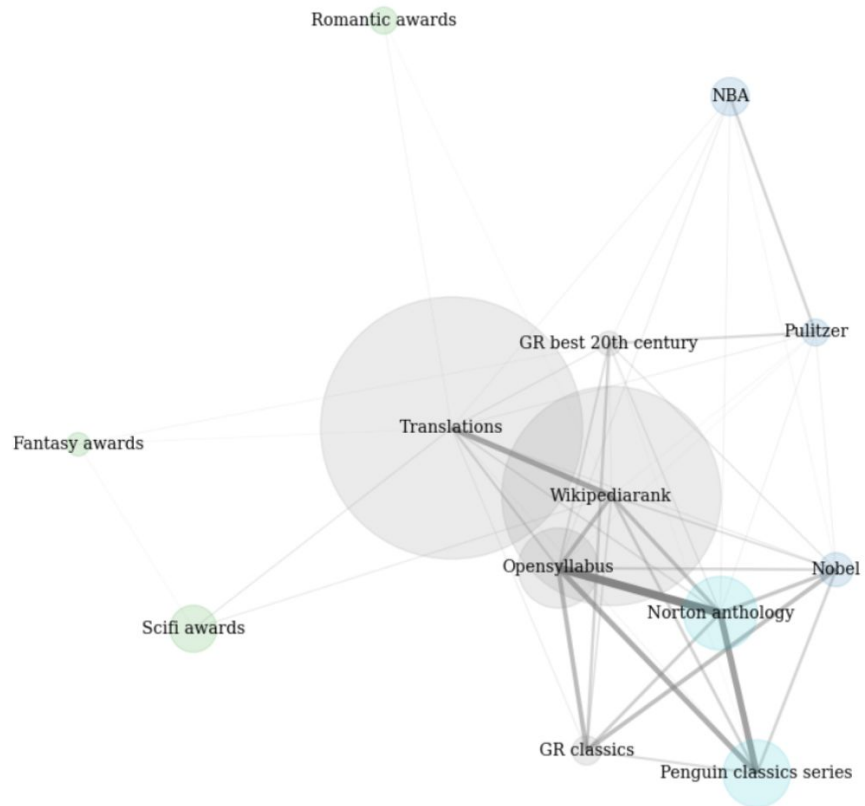


Award	Titles
National book award	108
Pulitzer prize	53
Nobel prize*	85
Scifi awards	163
Hugo award	
Nebula award	
Philip K. Dick award	
J.W. Campbell award	
Prometheus award	
Locus sci-fi award	
Fantasy awards	40
World fantasy award	
Locus fantasy award	
British fantasy award	
Mythopoeic award	
Horror awards	19
Bram Stoker award	
Locus horror award	
Romantic awards*	54
Rita awards*	
RNa awards*	

# Quality Metrics

## Hybrid Metrics:

- Number of translations
- Library holdings
- Crowdsourced lists





# Textual Metrics

## Readability and stylometry

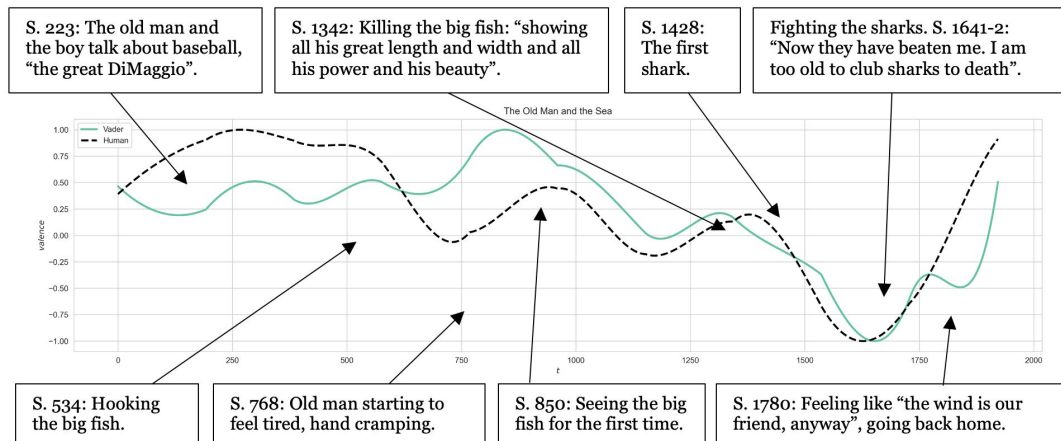
- Five different readability metrics
- Sentence length, word length and linguistic constants
- Most robust: Dale-Chall, includes a list a “easy words”
- Type-token ratio
- Compressibility
- Unigram and bigram entropy

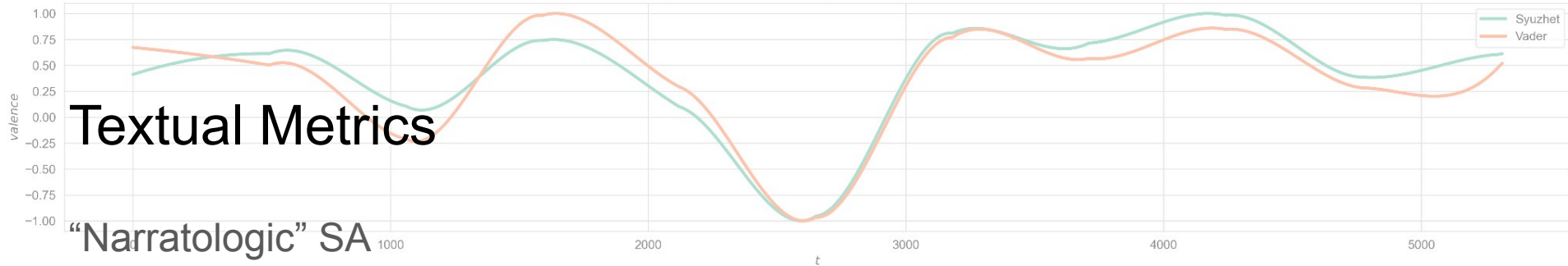
<b>Stylometric features</b>
Lexical diversity
Text compressibility (bzip2)
Sentence length
Avg. wordlength
Uni- and bigram entropy
LLM perplexity
<b>Readability measures</b>
Flesch reading ease
Flesch-Kincaid grade level
SMOG Readability formula
Automated readability index
New Dale-Chall readability formula
<b>Syntactic features</b>
Verb frequency
Noun frequency
Adjective frequency
Stopword frequency, etc.
<b>Syntax derived features</b>
Nominal/verb ratio
Percentage of passive verbs
Frequency of “of”, etc.
<b>Simple sentiment features</b>
Mean sentiment
Std. deviation sentiment
Ending sentiment
Beginning sentiment
Difference between main and ending
<b>Emotion features</b>
Emotion intensities
Emotion arcs
Emotion arc slope, etc.
<b>Sentiment arc measures</b>
Hurst exponent
Approximate entropy
<b>Metadata</b>
Author gender
Genre (manual annotation of subset)

# Textual Metrics

## Sentiment Analysis

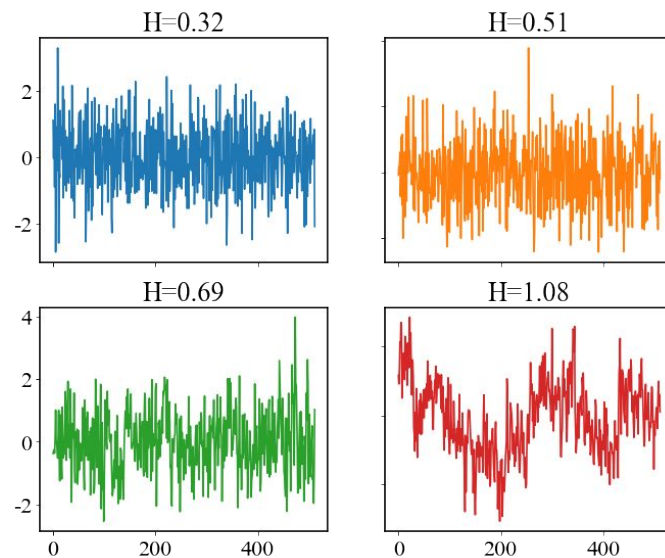
- Mean sentiment and standard deviation
- Sentiment arcs
- VADER and Syuzhet





- Hurst Exponent
- Approximate Entropy
- Arc Compressibility

	Mean	SD
Mean Sent.	0.03	0.04
Std Sent.	0.35	0.04
End Sent.	0.03	0.07
Beginning Sent.	0.04	0.05
Diff. Ending/Rest	0.01	0.05
Hurst Exponent	0.61	0.04
Approximate Entropy	1.75	0.15



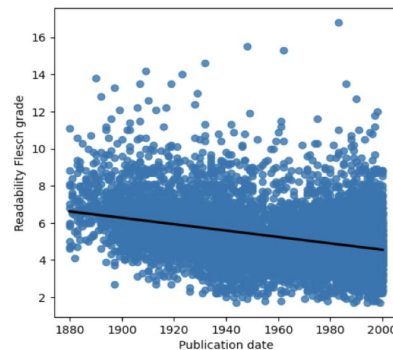
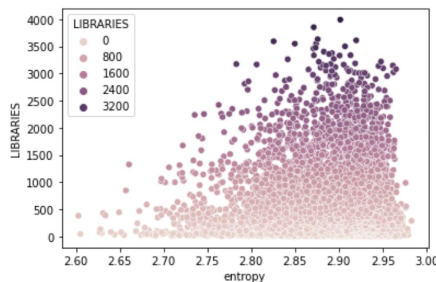
# Metadata

- Author, Title, Author's Gender, BookID
- Publishing Date and Location
- Genre Tags

Title	Auth_first	Auth_last	Publ	...	Avg_rating	SciFi Pr.	Pulitzer	Translations	...	Perplexity	Mean_sent	Readability
<i>A Clash of Kings</i>	George R. R.	Martin	1999	...	4.41	1	0	38	...	79.97	-0.002	92.73
<i>Dune</i>	Frank	Herbert	1965	...	4.25	1	0	398	...	72.74	-0.007	85.18
<i>Beloved</i>	Toni	Morrison	1987	...	3.92	0	1	68	...	68.78	0.030	91.71
<i>Misery</i>	Stephen	King	1987	...	4.20	0	0	74	...	68.09	-0.032	82.54
<i>The Portrait of a Lady</i>	Henry	James	1881	...	3.78	0	0	53	...	80.35	0.150	71.65

# Works using the resource

- Some works have already used the presented resource to explore the question of literary quality
- Train classifiers to gauge the perceived quality of unseen texts (Bizzoni et al. 2023)
- Explore the relation of quality proxies to textual characteristics (Wu et al. 2024, Feldkamp et al. 2024)



# Conclusions and Future Works

- Already allowed us to perform a number of unprecedented studies
- ***Many more features to come!***
- We hope to encourage quantitative analysis of reception and literary quality

# Thank You!

