A Statistical Look into the Occurences of the Anthropocene in Cli-Fi

Hana El Mourad

Table of contents

List of Figures

List of Tables

1 Part I: Introduction

1.1 Study Background

Literary works have long been considered a witness to the time they were conceived. Works of fiction depict social phenomena and capture a portrait of humanity's struggles. With that, every year, decade, and century brings new themes. From Greek Mythology and Roman Epics to Modernist and contemporary literature passing through Romanticism, Realism, and Naturalism, novels contain clues and signs of each age and reflect the obsessions and fears of each respective generation.

> From 2000, a group of geologists, led by the Nobel Prize winner Paul Crutzen, began to argue the present period of Earth's history should be known as the Anthropocene. […] According to proponents of the term Anthropocene, human activity has so altered the history of the Earth that it has become necessary to declare a new epoch to signify this impact. Trexler [2015, 1]

After the Second World War, a new geological era started, and with it, a new time of literature emerged: the literature of the Anthropocene or Post-1945 Fiction. These works of fiction carried in their pages reflections of the current state of planet Earth. Their scenarios often revolve around the end of the world due to human interference. This specific genre has been named Cli-Fi or Climate Fiction, a sub-genre of science fiction that delves into the intricacies of the climate crisis.

One of the main and most controversial books written on the topic of Anthropocene fiction is written by Amitav Ghosh, an Indian novelist, and literary critic**.** In his book, *The Great Derangement: Climate Change and the Unthinkable*, Ghosh accuses the modern novel writer of climate silence on one hand, and fetishizing climate disasters and setting them in such exaggerated forms that people do not take the events that occur in those novels seriously on the other hand. He says "There are surely very few writers today who are oblivious to the current disturbances in climate systems the world over. Yet, it is a striking fact that when novelists do choose to write about climate change it is almost always outside of fiction." Ghosh [2017, 8]

Other literary critics believe that the novel represents the perfect vehicle to warn people about possible futures if our human actions do not change. Such critics include Adam Trexler who published a book titled *Anthropocene Fictions: The Novel in a Time of Climate Change* in which he insists that "Climate novels have a role to play in our collective accounting of the Anthropocene, even those that were written in the hope that such a day would never come to pass." Trexler [2015, 237]

> The corpus of climate change fiction represents an enormous and growing archive. There cannot be a single history of these works, based on a corresponding history of climate science or climate politics, because the novels in question have continually reconfigured what it means to live in the Anthropocene, to be part of a self-conscious species that is actively transforming conditions that have reigned throughout its evolutionary history. Trexler [2015, 27]

In my Master thesis that I have written for my Master of Western Literature titled *In Defense of Science Fiction: Can Sci-Fi Novels Save the World*, I argued that Cli-Fi is an extremely powerful tool that can appeal to readers from every age and demography. Amitav Ghosh laments the current state of fiction and the silence surrounding the topic of "climate change" which he believes has resisted till this day the pages of serious fiction. While Ghosh laments the absence of climate change in serious fiction due to an "imaginative and cultural failure"Ghosh [2017]. In my thesis, I challenged Ghosh's views by putting 2 Sci-Fi novels under the spotlight and analyzing how each novel depicts the end of the world scenarios and how that makes them innocent of Ghosh's accusations of Climate deafness.

Expanding on this foundation, this paper will encompass the evaluation of an 'Anthropocene' corpus containing the two previously analyzed novels in addition to five works of fiction chosen for being emblematic of the Anthropocene. It has 490708 tokens and 25285 types, giving us a type-token ratio of 0.05. My goal through this research is to investigate this theory by using the language R to conduct an exploratory analysis and a keyword analysis and examine the words used in the Anthropocene literature. Therefore, the aim of this study is twofold:1- examining how each novel depicts the Anthropocene, and 2- Identifying recurring themes shared among the novels, shedding light on the prevailing concerns of Cli-Fi writers.

This paper is divided into two parts: Part I of this paper contains information about the study background, literature review, and the general Anthropocene themes. I then move to

discuss the Methodology as well as corpus description where information about both target and reference corpora is presented. Part II contains a statistical exploration of individual novels from the Anthropocene corpus by looking at collocations while using different association measures to assess their strength. From there, a keyword analysis is conducted where the Anthrocorpus (target corpus) is examined against a Bestsellers corpus (reference corpus) comprising eight random Bestsellers. Finally, an interpretation of the results is presented, and concluding remarks are made.

By taking several different novels written by different authors, with different backgrounds, styles, and plots, and examining their overarching themes, this paper strives to uncover the main preoccupation of Cli-fi writers. It is therefore my hope to prove that fiction writers are continuously inviting us to rethink our relationship with nature and the planet we dwell on through their texts, subtext, and context.

1.2 Climate Fiction Themes

The Anthropocene novels commonly explore themes of environmental devastation, resource scarcity, mobility/migration, and societal collapse as a result of climate change. They often depict a future world in which the effects of climate change have become catastrophic and irreversible, leading to the collapse of human civilization and the destruction of the natural world.

One of the main themes in Cli-Fi is the idea of *environmental devastation*, which is often depicted through the portrayal of extreme weather events, sea-level rise, pandemics and other effects of climate change. Disaster narratives paint different worlds where we can see the potential consequences of human actions and technology on the environment.

Another common theme is *resource scarcity*, which is often depicted through the portrayal of food and water shortages, energy crises, and other effects of climate change. These scenarios show how climate change will affect the availability of resources and how this will impact human society.

The Anthropocene novel also frequently addresses issues of forced *migration and displacement.* Environmental disasters, rising sea levels, and the degradation of land can lead to the displacement of individuals and communities, which becomes a central theme in these novels.

Societal collapse represents another of the main themes of the Anthropocene novel, which is often depicted through the portrayal of war, anarchy, and other effects of climate change. This theme usually revolves around human societies in the face of extreme climate crises and involves stories about the collapse of civilization as we know it.

1.3 Methodology

1.3.1 Data Collection and Pre-processing

The corpus was collected by converting ePub files into simple text (.txt) files. A clean-up of both corpora's text files was conducted whereby extra elements such as the preface, author's note, and acknowledgments were removed to provide better results. A stop list was also created and used to remove the most common connection words in the English Language. Moreover, the names of the main characters of each novel were removed, as well as unique, hyper-specific, details of the books (e.g. city/country names, organization names).

1.3.2 Project Script

The script mentioned in this paper was written in R. R is a programming language that is particularly helpful when it comes to examining large corpora or datasets. Its power and popularity among data scientists and linguists lie in its rich implementation as well as the availability of libraries. Using R [R Core Team, 2022] for statistical analysis with support from four packages {tidyverse} for data manipulation and visualization [Wickham et al., 2019], {here} [Müller, 2020] for easy file referencing, and finally {kableExtra} [Zhu, 2021]for table manipulation, different computations and visualizations were achieved.

The '{mclm}' package `{mclm}` is developed as a "companion to the Methods in Corpus Linguistics course at the Advanced Master in Linguistics (KU Leuven)" [Speelman and Montes, 2022]. It allows users to conduct various analyses and apply different techniques in the field of corpus linguistics. In this study, `{mclm}` was used to run different computations and statistical operations in order to examine how contemporary authors depict the end of the world and its causes.

The project was uploaded to GitHub. It includes several folders and files including a ReadMe file that describes the contents of the GitHub repository as well as the script in one R document, a stop-list, a quarto document, and other files such as a style sheet.

1.3.3 Techniques

This paper makes use of two main techniques used in corpus linguistics: collocations and keyword analysis. In order to use either techniques word frequencies must be calculated. Word frequencies are a very direct way of quickly assessing a text and its themes. It is also a very powerful tool when used alongside different data association measures. In the `{mclm}` package, word frequency is represented by *freqlist* which is a function that builds the word frequency list from a corpus.

> A central question in text mining and natural language processing is how to quantify what a document is about. Can we do this by looking at the words that make up the document? One measure of how important a word may be is its term frequency (tf). Silge and Robinson [2017, 31]

According to Stefan Evert, Evert [2007, 4] Chair of Computational Corpus Linguistics in Osnabrück, Germany, "we define a collocation as a combination of two words that exhibit a tendency to occur near each other in natural language," By looking at the collocations of these words or which words occur most frequently next to them, we get a better understanding of the context those words occur in.

A keyword analysis is a text or group of text compared to a reference corpus. While a collocation analysis focuses on certain words and sees which words are attracted to the context of that word. These methods are among the oldest and most used in the field of linguistics and have been practiced long before computers. Their popularity is attributed to their systematic and efficient way of uncovering patterns, themes, and linguistic features within large bodies of text. Moreover, they help reveal intricate relationships between words and how frequently they appear together in a text corpus.

> For keyword analysis, the target context is a target (sub)corpus and the reference context is a reference (sub)corpus. For collocation analysis based on surface co-occurrences, the target context is the text surrounding the occurrences of the node term and the reference context is all the other text in the corpus. Speelman and Montes [2022]

Association measures are how we describe the strength of attraction between two words or a word and a text. They are used in both techniques implemented in this paper. When we compare how often words occur together versus how often they do not, we can gain a better understanding of the context in which these words occur and extract insights about the text itself. There are two main categories of association measures: effect size measures and strength of evidence measures. To get results that are loyal to the size of the sample I was analyzing, it is important to use a combination of both. Effect size measures are measures that look at the size of the effect and how strong the attraction is but do not consider the sample size. That is why it is important to combine them with statistical measures that are going to be more loyal to the size of the data. The strength of evidence measures are statistical measures that look at how much difference there actually is between the observed and expected value. If you have a strong effect, you do not need a lot of data. Conversely, if you have a lot of data, you don't need as strong of an effect.

The first step taken in the analysis was to zoom in on each individual novel of the Anthropocene corpus by looking at the most frequent words by utilizing the *freqlist* function. The next step was to look closely at those words and examine their larger context by looking at their collocations to gain a better understanding of how they occur in the texts.

From there on, a keyword analysis was performed to extract the "words of the Anthropocene". To this end, this project also makes use of a second corpus used as a reference corpus to be able to run a keyword analysis and extract Anthropocene-centered words and themes. This is done by selecting the words that are "attracted" to a specific corpus in comparison to a reference corpus by looking at association scores (assoc_scores()).

For both collocations and keywords analyses, the output of assoc_scores() was evaluated following three criteria:

- A **frequency** of 3 or higher in the target context:
- A **PMI score** of 1.1 or higher.
- A **signed** G^2 score of 4 or higher.

Thresholds were indicated for all three association measures to eliminate low-frequency words and flow attraction words.

1.4 Corpus Composition

1.4.1 Anthrocorpus: The Target Corpus

The Anthropocene Corpus (Anthrocorpus) is a collection of seven **Cli-Fi** novels that address the theme of the Anthropocene, a geological era marked by the significant impact of human activity on the Earth's ecosystems. It has 490708 tokens and 25285 types, giving us a type-token ratio of 0.05.The corpus includes the following novels:

1- *Annihilation* by American Author Jeff Vandermeer - published in 2014

2- *Memory of Water* by Finnish Novelist Emily Itaranta - published in 2012

3- *Oryx and Crake* by Canadian Novelist Margaret Atwood - published in 2003

4- *Severance* by Chinese American Novelist Ling Ma - published in 2018

5- *Station Eleven* by Canadian Emily St. John Mandel - published in 2014

6- *The Road* by American Writer Cormac McCarthy - published in 2006

7- *Atmospheric Disturbances* by Canadian Israeli Writer Rivka Galchen- Published in 2008

These novels were selected for their relevance to the theme of the Anthropocene and the way they depict the relationship between humans and the environment. The corpus period spans from 2003 to 2018. This corpus, though compact, carries enough variety and weight to provide a unique opportunity to examine the preoccupation of modern writers with the end of the world.

1.4.2 Bestsellers: The Reference Corpus

The reference corpus used for keyword analysis was assembled by randomly selecting 8 best-selling fiction and non-fiction books. 13 books were initially selected, then 5 books were removed due to their relevance to the Anthropocene.

1. *Absalom's Daughters* by Suzanne Feldman - published in 2016, Historical Fiction
2. *It Ends With Us* by Colleen Hoover - published in 2016, Contemporary Romance
3. *The Bulgari Connection* by Fay Weldon - published in 2000, Literary Fiction
4. *The Silver Sun* by Nancy Springer - published in 1977, Sci-Fi/Fantasy
5. *The Three-Body Problem* by Cixin Liu(writer) Ken Liu(Translator) - published in 2006, Speculative fiction
6. *Triple* by Ken Follett - published in 1979, Thriller
7. *Trust* by Hernan Diaz - published in 2022, Historical Fiction
8. *Yumi and the Nightmare Painter* by Brandon Sanderson - published in 2023, Epic Fantasy

This was to allow for a later keyword analysis which would underline the words that are more unique to the Anthropocene Corpus.

2 Part II: A Deep Look Into the Anthropocene

2.1 Themes by Novel

In the following section, for the purpose of brevity in this paper, I will examine closely four of the seven novels and how each of those novels imagines the end of the world and how the themes of the Anthropocene manifest in those novels. As mentioned above, those themes include environmental devastation, resource scarcity, mobility/migration, and societal collapse. In order to mine the texts for those themes, I will resort to different methods and techniques used in text mining.

In the following sections, I will zoom in on each novel and examine closely the most frequent words, their collocations, and their association scores filtered by different strength measures.

Thus, word frequencies for 4 books from the Anthropocene corpus were computed. This allows us to explore immediately yet superficially words that tie our target corpus to the themes of the Anthropocene. However, only looking at how frequently a word occurs in a text is not fully indicative of the nature of the text. Context is also very important to look at, as well as statistical likelihoods. So, the next step was to select several words chosen out of the top 50

Table 1: The Road Most Frequent Words

rank	type	abs_freq	nrm_freq
1	road	256	43.59896
2	looked	229	39.00063
3	stood	201	34.23199
4	sat	180	30.65552
5	fire	132	22.48071
6	cart	125	21.28855
7	long	124	21.11825
8	water	109	18.56362
9	hand	108	18.39331
10	put	107	18.22300

most frequent words that could be relevant to the Anthropocene and its themes. I then looked at their collocation frequencies and their strengths of association, or which words appear more than others alongside the most frequent words.

2.1.1 The Road

The Road is a well-celebrated novel set in a post-apocalyptic world devastated by an unspecified disaster. The unknown catastrophic event leads to the collapse of human society and the destruction of the natural world. I will first take a look at the top most frequent words in this novel.

At first glance at Table 1, in the top 20 most frequent words we can see words related to mobility: *road, cart, walked.* We also see words such as *cold, dark, fire, blankets,* and *gray.* These words express the climate conditions the main characters live in. Ranking 10th, *water* and the theme of water are present in the top 20 most frequent words. *Dead* is the 18th most frequent word in the book.

From word frequencies, I then moved to look at association scores of select words that might be indicative of the Anthropocene. In *The Road*, I looked at the strength of association scores for the word *water* and its collocates.

Water scores 10th on the top frequencies list - it also has a lot of strong associations Table 2, and while they are mostly normal(drink, poured, glass) it could indicate a lack of water, as it is so prominently mentioned, and adjectives like heated, plastic, fresh, black, and gray could indicate there were issues with water scarcity or purity.

Table 2: Words with the highest association scores with the word Water

type	a	b	c	d	dir
heated	4.5	650.5	0.5	57954.5	1
drink	6.0	648.0	13.0	57941.0	1
drip	3.5	651.5	0.5	57954.5	1
poured	5.0	649.0	15.0	57939.0	1
bottle	5.0	649.0	20.0	57934.0	1
slow	4.0	650.0	12.0	57942.0	1
fresh	3.0	651.0	6.0	57948.0	1
glass	4.0	650.0	20.0	57934.0	1
plastic	5.0	649.0	58.0	57896.0	1
jars	3.0	651.0	15.0	57939.0	1
roadside	3.0	651.0	19.0	57935.0	1
watched	3.0	651.0	56.0	57898.0	1
good	3.0	651.0	64.0	57890.0	1
black	3.0	651.0	65.0	57889.0	1
gray	3.0	651.0	78.0	57876.0	1
sat	4.0	650.0	176.0	57778.0	1

Below is a plot visualizing the words with the highest association scores with the word *water* filtered by **signed** G^2 and measured against *PMI.* The size of the dots is indicative of frequency. We observe that the higher the *PMI* and **signed** G^2, the more they portray scarcity.

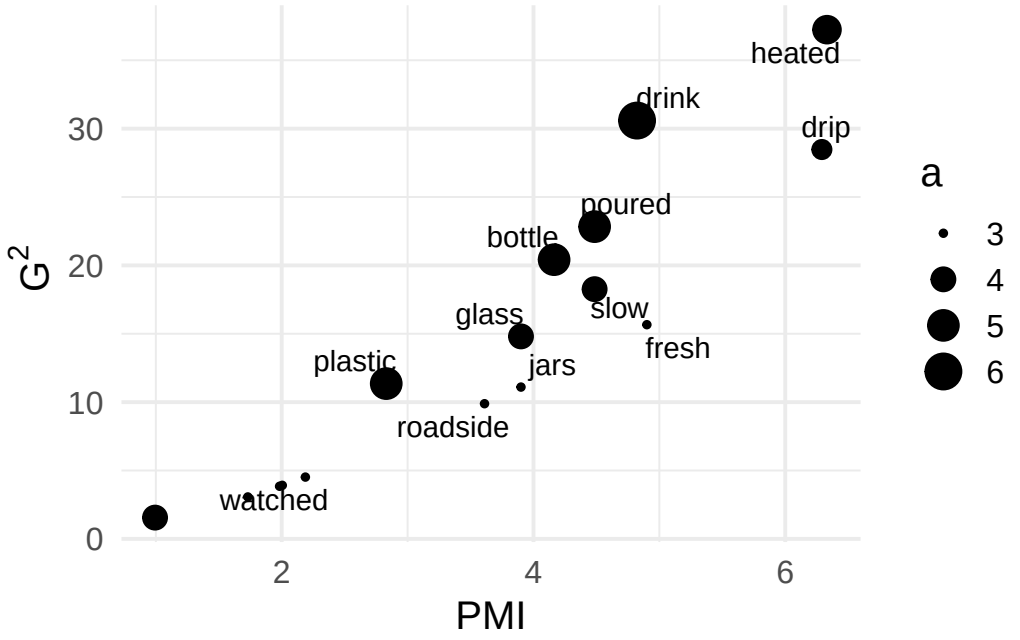

Figure 1: Association Scores For the Word Water in McCarthy

As seen in Figure 1, the words with the highest **signed** G^2 and **PMI** are *heated, drip, drink, poured, slow, fresh,* and *bottle. Heated* refers to the lack of hot water, lack of electricity, and infrastructure. *Drip* indicates slow stream of water or scarcity of water.

2.1.1.1 Co-occurrences of *Road in* McCarthy

As seen in Table 3, *road* has more than normal amount of strong associations, as is thematic to the book. e.g. side, crossed, stopped, stood, ran, south, ate, tracks, country.

I also looked at *Dead.* Following the analysis above, *Dead* has a strong attraction to words such as *trees, grass, limbs, black, wind,* and *fire.* These indicate the destruction of nature.

2.1.2 Severance

The novel is set in the aftermath of a pandemic and explores the theme of how humanity's actions have led to ecological disaster and the end of human civilization. The novel also critiques consumer culture, which is portrayed as a driving force behind the collapse of society.

Table 3: Words with the highest association scores with the word Road

type	a	b	c	d	dir
side	12	1515	56	56878	1
crossed	9	1518	34	56900	1
shuffling	3	1524	2	56932	1
curve	3	1524	3	56931	1
trees	7	1520	49	56885	1
watching	6	1521	44	56890	1
stood	13	1514	188	56746	1
stopped	7	1520	66	56868	1
ran	3	1524	13	56921	1
south	3	1524	18	56916	1
sat	10	1517	170	56764	1
ate	4	1523	39	56895	1
edge	3	1524	25	56909	1
tracks	3	1524	25	56909	1
country	3	1524	28	56906	1
looked	11	1516	218	56716	1
black	4	1523	64	56870	1
holding	3	1524	45	56889	1
standing	3	1524	46	56888	1
find	3	1524	53	56881	1
watched	3	1524	56	56878	1

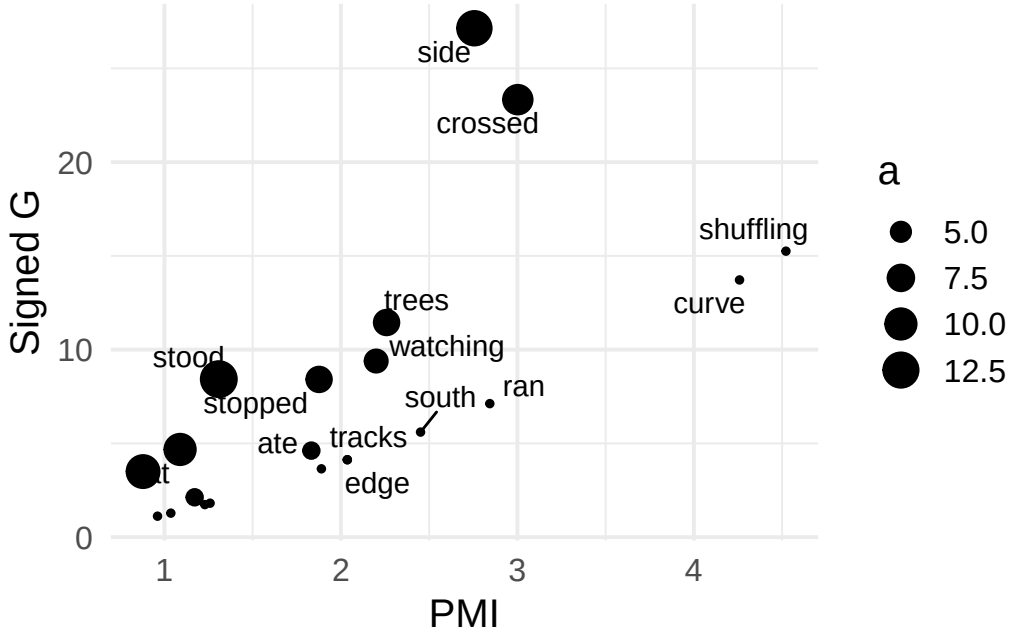

Figure 2: Association Scores For the Word Road in McCarthy

Table 4: Severance's Most Frequent Words

rank	type	abs_freq	nrm_freq
1	time	161	18.924034
2	looked	156	18.336331
3	office	108	12.694383
4	work	107	12.576842
5	night	97	11.401436
6	room	94	11.048815
7	day	83	9.755868
8	long	82	9.638328
9	place	81	9.520787
10	left	80	9.403246

Table 5: Words with the highest association scores with the word Face

type	a	PMI	G_signed	exp_a	dir
masks	5	5.707359	31.133455	0.0956938	1
wash	3	5.832890	19.261288	0.0526316	1
water	4	3.849378	14.180241	0.2775120	1
put	3	3.385431	8.797163	0.2870813	1
looked	3	2.006919	3.884822	0.7464115	1

In Table 4 *Office (109), work(107), and city(73)* all hint at capitalism and the metropolitan nature of New York City, the place in which this novel is set. In severance, Capitalism and hyper-globalization of the world have led to the rapid spread of the pandemic and the consequent downfall of humanity.

Fever occurs 59 times and *fevered* occurs 45 times both appearing in the top 50 most frequent words in Ma's book painting a very bleak image of the pandemic. I also observed that *city* (ranked 14th) has strong associations with 3 words: *live, leaving,* and *windows. Leaving* indicates migration from the city to safer pastures. *Windows* refer to the isolation during the pandemic. While *face* (ranked 16th has the most and strongest associations with *masks* with a **PMI** of 5.707 and **signed** G^2 of 31.133.

Another way to gain a better understanding of how the most frequent words occur in the text, in other words, their context, is to look at concordances (conc ()). Using this function allows us to look for a specific pattern using a regular expression (regex) and returns the concordance - which is the way this pattern or in our case, the most frequent word appears as well as the words that occur to the left and right.

> In the language sciences, concordancing refers to the extraction of words from a given text or texts (Lindquist 2009, 5). Commonly, concordances are displayed in the form of keyword-in-context displays (KWICs) where the search term is shown in context, i.e. with preceding and following words. Concordancing are central to analyses of text and they often represent the first step in more sophisticated analyses of language data (Stefanowitsch 2020).

Next, I zoomed in on the word *mask* in *Severance,* which is in the top most frequent word occurring 20 times in the text. This word has also the highest number of associations with the word *face.* In Table 5, we see *mask* appears next to *respirator, N95, epidemic, fever, fevered,* and others.These associations all fall under the theme of pandemics

```
Concordance-based data frame (number of observations: 20)
idx                                left|match|right
  1 ...ide each employee with two N95|masks|. If we wanted more, we could...
```

```
 2 ...here were two sets of N95 face|masks|and latex gloves, each imprin...
 3 ...rse weren't visible behind our|masks|. Only the stiffened cheeks. ...
 4 ...and Blythe took off their face|masks|. Blythe said, Should we just...
 5 ...ny policy for a reason. If the|masks|actually work, don't you thin...
 6 ...tive tools, such as gloves and|masks|to use when handling prototyp...
 7 ...eek, our voices muffled by the|masks|, in elevators in the morning...
 8 .... Our voices, amplified by the|masks|, sounded deeper and more gri...
 9 ...d begun to understand that the|masks|were not fever-preventative),...
10 ...d flocks, wearing useless face|masks|printed with I  NY. It was a...
11 ..., wearing a variety of fashion|masks|, in all black or leopard pri...
12 ...on magnified by the respirator|masks|that they'd put on as a joke....
13 ...models down the runway in face|masks|, gloves, and even scrubs, ma...
14 ...y by wearing those hideous N95|masks|, who was taking the initiati...
15 ...ned with the Supreme logo. The|masks|seemed to preclude any conver...
16 ... for all employees to wear N95|masks|in the office (before, this h...
17 ...Blythe donned their respirator|masks|, making jokes about "epidemi...
18 ...rs chanting, not wearing their|masks|so their voices could be hear...
19 ... he yelled. We put on our face|masks|and rubber gloves. We went in...
20 ...h Sentinel guards did not wear|masks|(given the scope of the epide...

This data frame has 6 columns:
   column
1 glob_id
2      id
3  source
4    left
5   match
6   right
```

2.1.3 Memory of water

The novel is set in a post-apocalyptic world where a pandemic has wiped out most of humanity. Droughts, wars, and water crimes are the result of humanity's attempts to control and manipulate nature through technology. Looking at Itaratna's top ten most frequently occurring words, we see the most frequent word by large margin occurring 131 in the text. *Plastic* is the fifth most occurring word in the novel. We also see *insect* and *past-world*.

Water scarcity is one of the main topics in Itaranta's book as seen in Table 6 . *Plastic* refers to the remnants of past society. While *insect* is part of a bi-gram *insect hood* which is a protective mask that people in *Memory of Water* wear to protect themselves from the insects that have taken over their landscape.

Table 6: Memory of Water Most Frequent Words

rank	type	abs_freq	nrm_freq
1	water	131	59.81462
2	house	56	25.56961
3	time	53	24.19981
4	face	40	18.26401
5	plastic	40	18.26401
6	turned	36	16.43761
7	inside	34	15.52441
8	insect	32	14.61120
9	past-world	31	14.15460
10	day	30	13.69800

Table 7: Plastic assocs

type	a	PMI	G_signed	exp_a	dir
grave	13	6.235196	103.09698	0.1725688	1
junk	3	6.119719	22.73728	0.0431422	1
metal	3	3.949794	11.31573	0.1941399	1

The word *plastic* as seen in Table 7 strong associations with *grave, junk,* and *metal.* This refers to the Anthropocene landscape that is covered with residues from the *past-world.*

I looked at *past-world* which occurs 30 times in the text and computed its concordances. We see it appears next words such as *books, echoes, era, technology,* and *winters.* These indicate nostalgia for older times as well as the destruction of human progress and societal collapse.

```
Concordance-based data frame (number of observations: 30)
idx                            left|  match   |right
  1 ...immer, and I longed for the|past-world|I had never known. I pictur...
  2 ...think about them, but their|past-world|bleeds into our present-wor...
  3 ... years old, I had read in a|past-world|book about snow and ice, an...
  4 ...ut them?' 'Why weren't more|past-world|books preserved?' I knew th...
  5 ...s drowned, why weren't more|past-world|books rescued?' 'I don't kn...
  6 ..., rain and sun had worn the|past-world|echoes thin a long time ago...
  7 ...village: constructed in the|past-world|era and converted later for...
  8 ...er all. Near the end of the|past-world|era the globe had warmed an...
  9 ...fused relic of the original|past-world|forms that have been long f...
 10 ...vely clear idea of what the|past-world|had been like - or rather, ...
 11 ...Sanja's home was one of the|past-world|houses, a one-storey with m...
 12 ...arrying the memories of the|past-world|locked within, slowly givin...
 13 ...' he said. 'I've heard of a|past-world|master who ordered his son ...
 14 ... seemed to run out, because|past-world|plastic took centuries to d...
 15 ... you find it weird how much|past-world|technology there still is i...
 16 ... always said at school that|past-world|technology was frail and ca...
 17 ...th this a major part of the|past-world|technology was gradually lo...
 18 ... was a random collection of|past-world|things excavated from the p...
 19 ...tinued to turn, but now the|past-world|voice was irrevocably gone....
 20 ... of the price of paper, and|past-world|volumes were virtually impo...
 21 ...es and summoned the feel of|past-world|winters about which I had r...
 22 ...arm, and I was no closer to|past-world|winters. I couldn't imagine...
 23 ...ey had been used for in the|past-world|, and I had only kept them ...
 24 ...aps then I'd understand the|past-world|, and the people who threw ...
 25 ...he broken technology of the|past-world|, metal and plastic intertw...
 26 ... comparison. One showed the|past-world|, the world of cold winters...
 27 ...istened to the voice of the|past-world|. At times it would wither ...
 28 ... little was known about the|past-world|. For all my winter daydrea...
 29 ...how winters had been in the|past-world|. I knew the darkness: ever...
 30 ... that the disc was from the|past-world|. We had been wrong. 'It's ...

This data frame has 6 columns:
   column
1 glob_id
```

Table 8: Station Eleven Most Frequent Words

rank	type	abs_freq	nrm_freq
1	time	197	20.31829
2	years	152	15.67705
3	road	150	15.47078
4	left	128	13.20173
5	night	126	12.99545
6	remember	117	12.06720
7	world	111	11.44837
8	room	107	11.03582
9	city	104	10.72640
10	light	103	10.62327

```
2      id
3  source
4    left
5   match
6   right
```

2.1.4 Station Eleven

The novel is set in the aftermath of a pandemic and explores the theme of how humanity's actions have led to ecological disaster and the end of human civilization. The novel also critiques consumer culture, which is portrayed as a driving force behind the collapse of society.

road and left represent the theme of mobility; *water* represents water scarcity. *years, time, and remember* represent the nostalgia to pre-apocalypse times.

The word *pandemic* is observed 16 times in *Station Eleven.* When looking at its concordances, we see it appearing next to words such as *virus* and *flu* which give us clues into the nature of the pandemic. We also see *has started, hit, reach, happened,* and *reach* which detail the spread of the disease in this novel.

```
Concordance-based data frame (number of observations: 16)
idx                            left| match  |right
  1 ... one where there had been no|pandemic|and he'd grown up to be a ph...
  2 ...sts, so no one mentioned the|pandemic|as she crossed the lobby, al...
  3 ...r one where there had been a|pandemic|but the virus had had a subt...
  4 ...tine theory, namely that the|pandemic|had started in Europe, the l...
```

```
  5 ...range ideas?" "He thinks the|pandemic|happened for a reason," Clar...
  6 ...tly obsolete by the time the|pandemic|hit, but used by a few peopl...
  7 .... August remembered his pre-|pandemic|life as an endless sequence ...
  8 ...course the context, the pre-|pandemic|world that he remembered so ...
  9 ...d a dozen shopping bags. The|pandemic|would reach North America in...
 10 ...ned to CNN. The story of the|pandemic|'s arrival in North America ...
 11 ...For a whole decade after the|pandemic|, I kept looking at the sky....
 12 ...emplating the flu, the great|pandemic|, and let me ask you this. H...
 13 ...n the face of a probable flu|pandemic|, but couldn't resist. He pl...
 14 ...hat it was developing into a|pandemic|. But now he watched the too...
 15 ...s was three weeks before the|pandemic|. They still had the indescr...
 16 ...ibly be out there except the|pandemic|? Nonetheless, the scouting ...

This data frame has 6 columns:
   column
1 glob_id
2      id
3  source
4    left
5   match
6   right
```

2.2 Statistical Analysis

After zooming in on each novel and examining how each novel addresses the impact of the Anthropocene in the previous section, the study zooms out and looks at the corpus as a whole. The following section covers the keyword analysis where I look at which words are attracted to the target corpus versus the reference target.

2.2.1 Keyword Analysis

In the keyword analysis, the Anthrocorpus(target) is compared to the bestsellers(reference) corpus, and by comparing how often words occur in the target corpus vs in the reference corpus, it is calculated whether a word is occurring more frequently in the target corpus, than it is expected to occur. By looking at which words are significantly more attracted to each other in target versus reference, we gain a better understanding of the context and the keywords that are distinctive to each text.

The strength of the results is further confirmed by the assoc_scores function, which is filtered by PMI which compares the observed frequency versus the expected frequency ($PMI >= 1.1$), the Signed G^2 strength measures($G\_signed >= 4$), as well as the frequencies (a > 3). *PMI* compares the observed frequency to the expected frequency. A *PMI* higher than one means

Table 9: Top 20 Keywords of the Anthrocorpus filtered by frequency, PMI, and G-signed.

type	a	PMI	G_signed
plastic	186	1.254662	251.34889
cart	152	1.311994	233.18285
snow	137	1.100698	133.94929
lighthouse	101	1.394969	195.30841
beach	99	1.191450	117.03721
expedition	81	1.291890	118.72513
prophet	81	1.377281	146.80631
fever	74	1.145005	79.39108
border	68	1.160518	75.35601
tarp	62	1.394960	119.88091
bible	56	1.179971	64.63021
flu	51	1.366969	89.70369
crawler	48	1.394953	92.80547
compound	47	1.168213	52.92684
fevered	47	1.364610	82.12948
mall	47	1.394953	90.87154
facility	45	1.332248	72.39810
photo	45	1.131949	46.97683
insect	41	1.260682	56.12118
paradice	41	1.394948	79.26800

the collocation occurs 2 times more often than expected. While Signed G^2 is a likelihood ratio test statistic also called the G-test of goodness-of-fit.

As previously discussed, combining different effect size measures and strength of evidence allows for a clearer view of how the Anthropocene words are used and offers a more solid grasp of the context whether it is the text of each novel or the Anthropocene corpusas a whole. By doing so, we are able to look at the strength of the effect size measures which adds to the level of certainty that there is indeed a difference between the target corpus and the reference corpus. Moreover, we get to zoom in on the Anthropocene-specific words and examine their themes.

The keyword analysis returns 1150 types or unique words in the list. An exploration shows that words such as *tarp, flu, fevered, carts, emigrant, masks, spores, past-world, sample expeditions, caravans, trailer, virus, and can* - stand out. In the table below, we looked at the top 100 most frequent words in the Anthrocorpus and then we categorized them according to their relevant

themes. This shows that those are some of the words somewhat typical to books written about the Anthropocene/Cli-Fi.

Table 10: Words of the Anthropocene categorized by themes

Anthropocene Theme	Word
Religion	Prophet, baptists, bibles, paradice, ark, congregation, devotees
pandemics	flu, fevered, crawler, pesthouse, masks, spores, virus, corpse, fungal, pandemic, quarantine, vaccine, parasites, microbes
movement and migration	emigrants, caravans, ferry, trailer, carts, camper, carriage, migrants
nostalgia	past-world, iPhone, mall, walmart, googled, facility, compound, photo,
food scarcity	canned
water crisis	tidewater, water, waterskin

2.3 Interpretation

The various analyses of word frequencies, collocates, and keywords carried on in this study returned interesting results about the prevalent themes and contextual nuances of the Cli-Fi novels.

Through a thorough exploration of the word frequencies, we saw certain terms recur frequently within the corpus, reflecting key concerns of the Anthropocene genre. In McCarthy's book, words like *road, water,* and *dead* dominate the most frequent word list, indicating a focus on mobility, resource scarcity, and post-apocalyptic uninhabitable landscape. In Ling Ma's *Severance,* we get to see themes of consumer culture contrasted with societal collapse through words such as *office, work,* and *city.* We encountered pandemic-related terms in both *Severance* and *Station Eleven* such as *fevered, masks, contagious, etc.*

Furthermore, the analysis of collocations provides a deeper understanding of how specific words are semantically linked within the context of each novel. For instance, in *The Road,* the collocations of the word *water* reveal associations with words like *heated, drip,* and *drink,* suggesting the struggle for water access in a desolate landscape. Similarly, the collocations of *dead* in the same novel show connections with terms like *trees, grass,* and *fire*, depicting vivid imagery of environmental devastation and loss. Such collocation patterns expose the thematic richness and narrative intricacies that amplify the Anthropocene discourse in Cli-Fi literature.

In the keyword analysis, a comparative study between the Anthropocene target corpus and the reference corpus of bestsellers provides quantitative evidence of distinctive words and themes in the former. The presence of words like *flu, fevered, masks, and pandemic* underscores the emphasis on pandemics and health crises, which resonate with the contemporary climate of

global concerns. Additionally, terms such as *emigrants, caravans, and trailer* point towards the recurring motif of movement and migration, reflecting the displacement and mobility resulting from climate-induced challenges. This keyword analysis reaffirms the thematic trends observed in the word frequencies and collocation analyses, highlighting the presence of recurring concerns in Anthropocene literature, from resource scarcity to ecological deterioration and societal upheaval.

In conclusion, the use of word frequencies, collocation patterns, and keyword associations in this study provides a comprehensive lens through which the themes of the Anthropocene are examined. These analyses collectively illustrate the multidimensional nature of Anthropocene literature, enabling a nuanced exploration of the complex interplay between humanity, the environment, and the consequences of our actions. Through the examination of textual patterns, this study highlights the narrative strategies employed by authors to engage readers in critical conversations surrounding climate change, resource exploitation, and the fragile balance between human civilization and the natural world.

Reflecting on the results of this study, we can see that all seven novels carried the Anthropocene themes in their lines. The themes of migration/displacement, food and water scarcity, and post-apocalyptic landscape are strongly present in each of the novels.

Moreover, when looking at the concordances of Anthropocene-specific terms, we were able to see in juxtaposition the state of the earth before and after human intervention has destroyed the planet rendering it unlivable.

2.4 Limitations

Though the size of both target and reference corpora is limited, the different analyses conducted and the combination of the effect size and strength of evidence association measures resulted in a list of Anthropocene words thus answering our research question. The corpus, while carefully selected, contains only seven novels, which may limit the generalization of the findings. Therefore, it would be curious to inspect a corpus with a larger scope, sample size and composition.

Aside from the size of the data, another limitation would be the removal of certain words, names, or hyper-specific details from the text that could potentially have affected the context and meaning of the analyzed passages. Decisions about what to exclude and what to include could have introduced unintended biases.

Additionally, while the study used a combination of statistical and effect size measures, it could benefit from advanced statistical analysis such as linear regression or correspondence analysis. These methods could provide stronger evidence and a wider understanding of the Anthropocene corpus.

In summary, this study provided valuable insights into the themes and linguistic features of the Anthropocene literature. However, the limitations listed above should be acknowledged to ensure a comprehensive and loyal interpretation of the study's findings.

3 Conclusions

The Anthropocene Novel has been the topic of literary focus in recent years due to the growing number of fiction that imagines the end of the world as propelled by humankind. This genre presents a valuable opportunity for understanding how literature deals with climate change and the different environmental issues that humanity is facing.

> When the novel incorporates things implicated in climate change—climate models, glaciers, cars, future hopes, weather—it becomes impossible to read without the preoccupation of climate change. […]These sites are integral to the meaning-making of a novel, and each of them is being radically reordered as we locate ourselves in the Anthropocene. […] As still more novels incorporate the weather, technology, and ideas of the Anthropocene, features of these early climate change novels will be diffused into literature at large. Trexler [2015, 16]

Therefore, the main task of science fiction is not to prophesize and predict the future, but rather explore the different ways the future may unfold**,** penetrate our collective conscience, and guide humanity to change its ways. By mining the text for keywords and looking at their larger context, examining their occurrences and co-occurrences within a given text, and then comparing the Anthrocorpus to a randomized literary corpus, this study in its two parts allowed us to gain a better understanding of the overarching themes of the Anthropocene at first glance.

The first part introduced the research question along with background information and the themes of the Anthropocene. I then move on to talk about the methodology followed in the analysis as well as the corpus composition. In the second part of the paper, we get a deeper look into the Anthropocene literature by first zooming in on the individual target corpus novels and their word frequencies. On top of that, I zoomed in closer into select words from the most frequent lists and looked at their concordances to gain a better understanding of how those words occur in the texts. Furthermore, I zoomed out to look at the Anthropocene as a whole and analyze it in comparison to a target corpus by running a keyword analysis that allowed us to extract Anthropocene-specific terms.

Trexler insists that "the imaginative capacities of the novel have made it a vital site for the articulation of the Anthropocene." We can add to that the linguistic nature and power of words and how they are used in the Anthropocene novel also play an essential role in shaping our collective consciousness and ultimately affect our actions and perspectives. Astronomer Carl Sagan believes in the power that science fiction holds through which it allows humanity to explore "alternative futures, both experimental and conceptual". To Sagan, "the greatest

human significance of science fiction may be as thought experiments, as attempts to minimize future shock, as contemplations of alternative destinies." In fact, it has been argued that the human ability to imagine, is necessary so that the imagined "person" dies, instead of the real one.

After careful examination of the recurrent themes of the Anthropocene through a thorough analysis of the different keywords and concordances that occur in the text, this paper was able to identify the words of Anthropocentric nature. The novels analyzed in this study portrayed the devastating effects of human activity on the planet and the consequent environmental disasters and paint a very vivid and eerily close image of the realities of the climate crises. Thus, the contemporary writer is considered innocent of Amitav Ghosh's accusation of climate deafness.

References

Stefan Evert. Corpora and collocations. *Institute of Cognitive Science, University of Osnabrück*, 2:1212–1248, 2007. doi: doi:10.1515/9783110213881.2.1212. URL http://www.jstatsoft.org.

Amitav Ghosh. *The Great Derangement: Climate Change and the Unthinkable*. The University of Chicago Press, 2017.

Kirill Müller. *here: A Simpler Way to Find Your Files*, 2020. URL https://CRAN.R-project.org/package=here. R package version 1.0.1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Julia Silge and David Robinson. *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc., 1st edition, 2017. ISBN 1491981652. URL https://www.tidytextmining.com/.

Dirk Speelman and Mariana Montes. *mclm: Mastering Corpus Linguistics Methods*, 2022. https://github.com/masterclm/mclm.

Adam Trexler. *Anthropocene fictions: The novel in a time of climate change*. University of Virginia Press, 2015. ISBN 0813936934.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Hao Zhu. *Construct complex table with "kable" and pipe syntax [R package kableextra version 1.3.4]*, Feb 2021. URL https://cran.r-project.org/web/packages/kableExtra/index.html.