Coffee break 1

Anusha Bhat and Aditi Mannem

October 3, 2023

1 Introduction

For this report, we aim to evaluate the differences between young adult books and children's literature books by investigating the most frequently used words, as well as the usage of pronouns, and mentions of body parts. Identifying which terms are used and how they are used allows one to understand the themes between the two genres. Our objective is to understand how identity and level of maturity evolve over childhood and adolescence in writing. Since literature is at the forefront of shaping how we think and perceive the world, it is important to highlight similarities and differences between the two genres to elucidate potential differences in how children and young adults view the world and act. We hypothesize that the young adult genre will have a higher usage of words with mature and complex meanings when compared to the children's literature genre due to the more explicit nature of the young adult genre.

2 Data

For our data, we will use two corpora–one specific to young adult books and the other specific to children's literature. The young adult corpus was generously provided to us via the "36-468: Text Analysis" Canvas page by Dr. David Brown. This corpus originally consisted of four book series, from which we randomly shortlisted two series for the sake of computational complexity. This corpus currently consists of the "Twilight" series by Stephanie Meyers and the "Divergent" Series by Veronica Roth. The Twilight series consists of the following novels: "Breaking Dawn", "New Moon", "Eclipse", and "Twilight". The Divergent series consists of the following novels: "Divergent", "Insurgent" and "Allegiant".

Token Counts and Percents by Text: Young Adults Corpus

	Tokens	Percent
Meyer_BreakingDawn.txt	175733	20.62795
Meyer_Eclipse.txt	141035	16.55502
Meyer_NewMoon.txt	125788	14.76529
Meyer_Twilight.txt	111967	13.14295
Roth_Allegiant.txt	101722	11.94037
Roth_Divergent.txt	97008	11.38702
Roth_Insurgent.txt	98664	11.58141

[1] Table 1

In table 1, we can observe the corpus composition by token count and percentage for each of the books. "Breaking Dawn" from the "Twilight" series composes the largest portion (20.63 %) of the corpus while "Insurgent" from the "Divergent" series composes the smallest portion (11.50 %). The corpus has a total of 851,917 tokens. The children's literature corpus was sourced from EDENDB's "Children Stories Text Corpus" dataset on Kaggle. The original corpus contains hundreds of books and short stories, which we downsized to fifty books via random selection. Additionally, we created fifty .txt files from the original single .txt file for easier analysis.

Sample Token Counts and Percents by Text: Children's Literature Corpus

	Tokens	Percent
the_snow_queen.txt	13385	8.839359
the_shoes_of_fortune.txt	11058	7.302625
the_yellow_dwarf.txt	6934	4.579165
beauty_and_the_beast.txt	6349	4.192835
thie_white_cat.txt	5363	3.541687
the_wonderful_sheep.txt	5251	3.467723
the_master_maid.txt	5219	3.446591
aladdin_and_the_wonderful_lamp.txt	4525	2.988278
the_shadow.txt	4351	2.873370
prince_darling.txt	4160	2.747235

[1] Table 2

In table 2, we can observe the corpus composition by token count and percentage for ten stories that made up the largest proportion of the corpus. "The Snow Queen" composes the largest portion of the corpus (8.84 %). The corpus has a total of 151,425 tokens.

To create our corpora, we first tokenized the datasets and removed numbers, punctuations, and symbols to ensure that only words remained. We also converted all the tokens to lowercase for consistency during our analysis. In addition, we removed the articles "a", "an", and "the", and the common conjunctions "for", "and", "nor", "yet", "so", "still", and "besides". This makes our analysis more meaningful since we removed highly common words used in modern language that do not really show any significance when looking at what tokens are the most frequent in our corpora. Furthermore, removing these words allows us to see what unique words are widely used in our corpora.

Along with our overarching corpora, we made additional sub-corpora to analyze the usage of pronouns and body parts in each genre–a body parts subcorpus and a pronouns subcorpus. Each original corpus has these two sub-corpora, for a total of four sub-corpora. We created the body parts sub-corpora by keeping the tokens corresponding to the following body parts: face, lips, eyes, mouth, nose, ears, fingers, hands, hair, skin, blood, bones, ear, finger, eye, faces, lip, hand, ankle, ankles, wrist, wrists, back, backs, stomach, foot, and feet. Finally, we created the pronouns subcorpora by selecting the tokens corresponding to the following pronouns: she, he, they, them, their, hers, his, theirs, him, I, you, it, we, us, me, mine, ours, its, your, and yours.

3 Methods

To begin our analysis, we first identified what tokens had the highest frequency of occurrence and how they were used in the original corpora. We then looked at the subcorpora and further identified which body parts and pronouns were most frequently used in each genre as well as what words they collocated with. This allowed us to examine overarching differences between themes and word usage in each genre. After this initial investigation, we looked at keyness values and effect sizes to identify if there are any words that have significantly different frequencies between the two genres. Finally, we concluded our research with a sentiment analysis to identify if there are any potential differences in emotional valence between the genres.

4 Results

4.1 Exploring the Young Adult Corpus

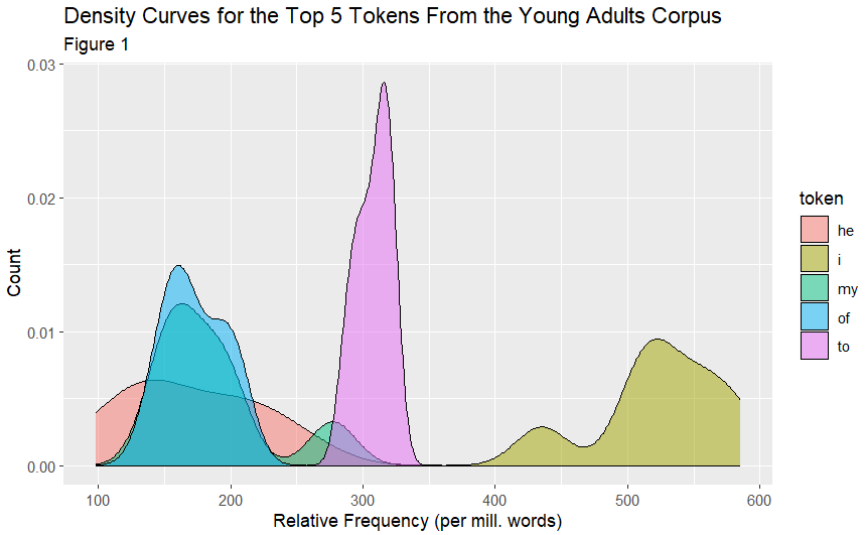
We begin our exploratory data analysis of the young adult corpus by examining what tokens occur the most frequently.

Most Frequent Tokens: Young Adults Corpus

feature	frequency	rank	docfreq	group	RelFreq
i	44109	1	7	all	51776.17
to	26173	2	7	all	30722.48
my	15505	3	7	all	18200.13
of	14499	4	7	all	17019.26
he	14492	5	7	all	17011.05

[1] Table 3

In table 3, we can see that the words "i", "to", "my", "of", and "he", are the top five most frequent tokens in the corpus, all of which appear in all seven of the novels. The most frequent token is "i" with a total frequency of 44,109 times and a relative frequency of 51,776.17. "He" has the lowest frequency among the top five tokens with a total frequency of 14,492 times and a relative frequency of 17,011.



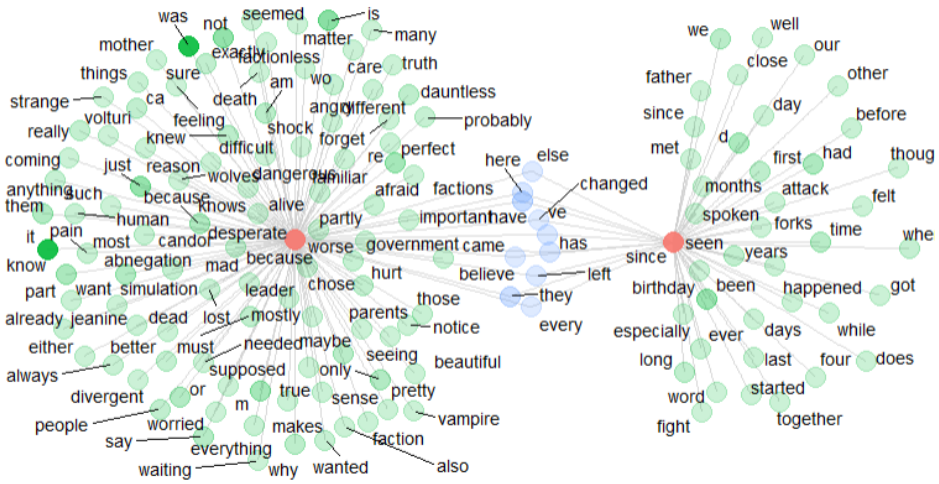
In figure 1, we observe that although "i" has the highest total relative frequency, "to" has the highest relative frequency for a single novel in the corpus. In fact,

the tokens "my" and "of" also have higher single-text relative frequencies than "i". From this, we notice that "i" may not be occurring as frequently in a single novel compared to the other top five tokens, however, it occurs more than the others when you account for all seven novels in the corpus.

Next, we created a collocation network for the words "because" and "since". These are words that are common in the vocabularies of English-speaking youth, so we wanted to inspect how they were incorporated into the young adult novels.

Figure 2

Collocation Network of "because" and "since"



In figure 2, we can observe that more words collocate with "because" than "since". The words "desperate", "alive", "partly", "worse", "mad", "lost", "leader", "chose", and "candor" have the strongest collocation association with "because" while the words "seen", "birthday", "spoken", "months", and "years" have the strongest collocation association with "since". Similarly, "was", "it", "not", "just", and "is", have the highest collocation frequencies with "because" meanwhile "been" and "had" have the highest collocation frequencies with "since". There are several words that collocate with both "because" and "since" such as "here", "else", "have", "changed", "has", "left", "came", "believe", "they", and "every".

4.2 Exploring the Children's Corpus

We begin our exploratory data analysis of the children's corpus by examining what tokens occur the most frequently in a similar manner as we did with the young adult corpus.

Most Frequent Tokens: Children's Corpus

feature	frequency	rank	docfreq	group	RelFreq
to	4758	1	50	all	31421.50
of	3394	2	50	all	22413.74
he	3254	3	49	all	21489.19
was	2789	4	50	all	18418.36
in	2414	5	50	all	15941.89

[1] Table 4

In table 4, we notice that the top five most frequent tokens in the children's corpus are "to", "of", "he", "was", and "in". The word "to" occurs the most frequently with a total frequency of 4,758 times and a relative frequency of 31,421.50 while "in" occurs the least frequently of the top five tokens with a frequency of 2,414 and a relative frequency of 15,941.89. Both the young adult and children's corpus have the words "to", "of", and "he" in their top five most frequent tokens.

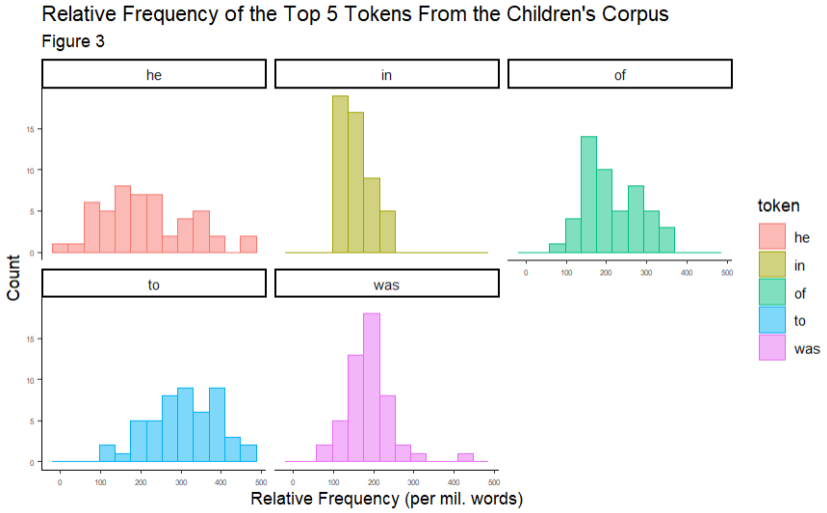
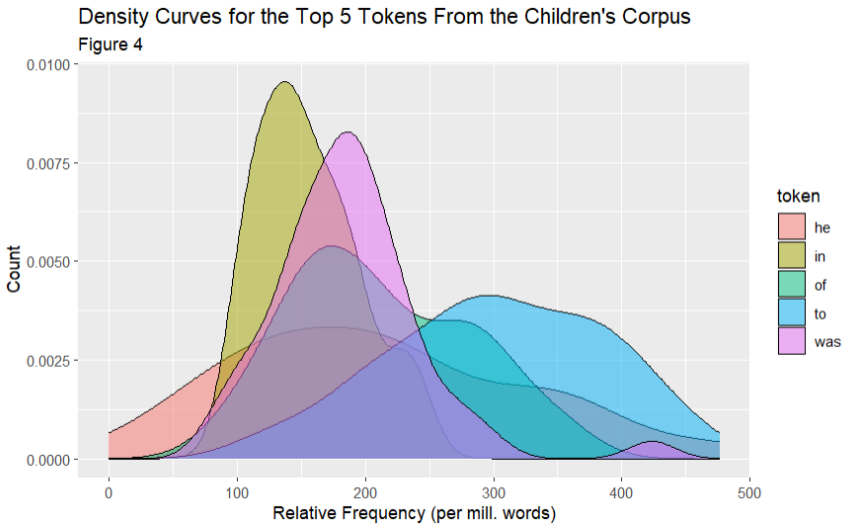


Figure 3

In figure 3, we display histograms faceted by the top five tokens to observe their distributions for relative frequency across the fifty short stories. The tokens "he", "of" and "to" all have unimodal and symmetric distributions that follow a bell curve shape, suggesting that these tokens occur at similar counts respective to their relative frequencies across each of the stories in the corpus. On the other hand, "in" and "was" have unimodal but nonsymmetric distributions that are slightly skewed, suggesting that there is an uneven distribution of these two tokens between the stories (e.g., some stories may have a higher frequency of "was" while some have a lower frequency).

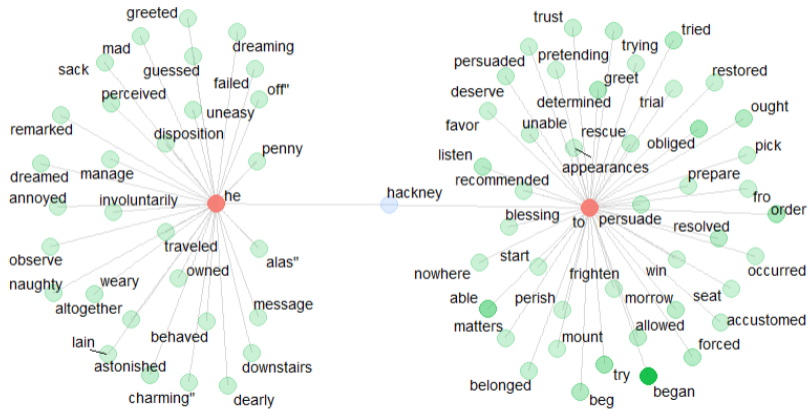


From figure 4, we can see that "in" has the highest single-text relative frequency while "to" has the lowest single-text relative frequency despite "to" having the highest relative frequency overall. This is the same phenomenon that we observed in a similar graph constructed for the top five tokens in the young adults corpus.

Next, we looked at collocations for "to" and "he" since "to" has the highest frequency and "he" was also a top five frequent word in the children's corpus.

Collocation Network of "to" and "he"

Figure 5



From figure 5, we noticed that the words "traveled", "owned", "alas", "penny", "disposition", "manage", and "involuntarily" have the strongest collocations with "he" while the words "persuade", "frighten", "start", "blessing", "recom-

mended", "prepare", and "appearances" all have the strongest collocation association with "to". "Began", "able", and "order" have the highest collocation frequency with "to", whereas, "astonished" and "perceived" had the highest collocation frequency with "he". Only "hackney" collocates with both "to" and "he". It appears that more verbs collocate with "to" meanwhile more feelings and descriptors collocate with "he".

4.3 Identifying Differences Between the Corpora

4.3.1 Looking at the Usage of Body Parts Terminology

Our first formal investigation of differences between the two genres begins with dividing our corpora into respective subcorpora that only include a select list of body parts terms.

Most Frequent Tokens: Young Adult Body Subcorpus

feature	frequency	rank	docfreq	group	RelFreq
eyes	2718	1	7	all	181563.13
back	2692	2	7	all	179826.32
face	2028	3	7	all	135470.94
hand	1497	4	7	all	100000.00
hands	1042	5	7	all	69605.88

[1] Table 5

Most Frequent Tokens: Children's Body Subcorpus

feature	frequency	rank	docfreq	group	RelFreq
back	233	1	37	all	224254.09
eyes	136	2	39	all	130895.09
hand	127	3	37	all	122232.92
feet	84	4	24	all	80846.97
hands	80	5	28	all	76997.11

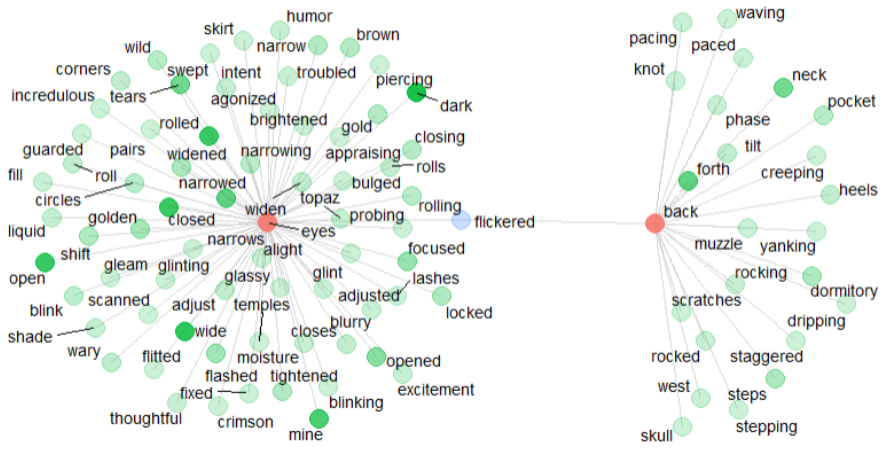
[1] Table 6

From tables 5 and 6, we note that "eyes", "back", "hand", and "hands" are among the top five most frequent body parts terms in both subcopora. "Hands" is the fifth most frequent for both of the corpora. One difference is that "face" is among the top five frequent tokens in the young adult body subcorpus, while "feet" is among the top five in the children's body subcorpus instead. Both "eyes" and "back" are the top two most frequent between the two subcorpora, although, they are ranked 1 and 2 respectively for the young adult body subcorpus but 2 and 1 respectively for the children's body subcorpus.

Next, we looked at the collocation of "back" and "eyes" in the original corpora since these were the top two most frequent tokens in the subcorpora. Our goal with these collocation networks was to potentially elucidate how these two words are being used in each genre.

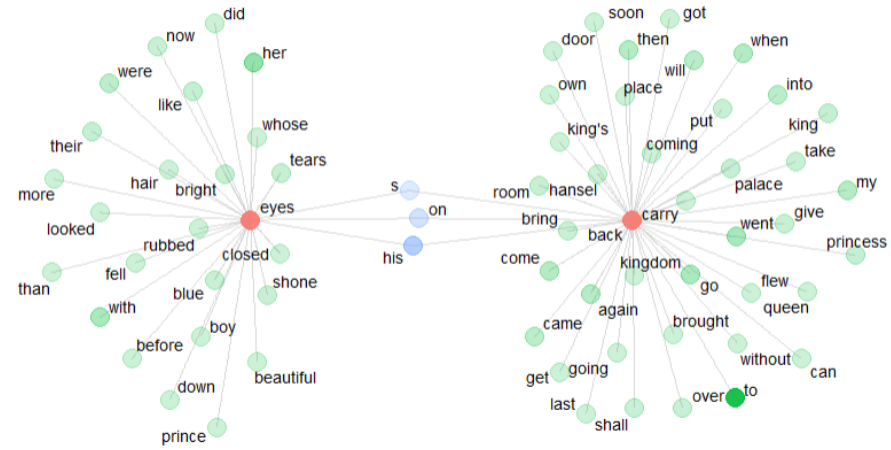
Collocation Network of "back" and "eyes": Young Adult Body Subcorpus

Figure 6



Collocation Network of "back" and "eyes": Children's Body Subcorpus

Figure 7



We see that between both corpora, more words collocated with "eyes" in the young adult corpus than in the children's corpus. In particular the to-

kens "rolled", "narrowed", "closed", "swept", "piercing", "widened", open", "wide", and "mine" collocated the most frequently with "eyes" in the young adult corpus. In the children's corpus, "his", "her", "with", collocated the most frequently with "eyes". We note that more descriptive words for physical attributes were collocated with "eyes" in the young adult corpus. This could perhaps be due to the more complex nature of the writing and the fact that novels are inherently more verbose and in-depth than short stories. It could also be due to more romance or other one-on-one scenes in young adult novels that would warrant a higher frequency of describing other characters.

On the other hand, we see that between both corpora, more words collocated with "back" in the children's corpus than in the young adult corpus. In the young adult corpus, "forth", "dormitory", and "neck" collocated the most frequently with "back". Meanwhile, in the children's corpus, "to", "his", "went", and "when" collocated the most frequently with "back". More violent and negative words collocated with "back" in the young adult corpus than in the children's corpus. This could be because there are many fight scenes in both Twilight and Divergence rather than a reflection of the young adult genre.

4.3.2 Looking at the Usage of Pronouns Terminology

For this section, we employed the same methods from section 4.3.1 but this time our subcorpora were based on a list of pronouns rather than body parts.

Most Frequent Tokens: Young Adult Pronouns Subcorpus

feature	frequency	rank	docfreq	group	RelFreq
i	44109	1	7	all	334392.15
he	14492	2	7	all	109864.45
it	13755	3	7	all	104277.22
you	12930	4	7	all	98022.86
me	10799	5	7	all	81867.67

[1] Table 7

Most Frequent Tokens: Children's Pronouns Subcorpus

feature	frequency	rank	docfreq	group	RelFreq
he	3254	1	49	all	186208.87
it	2053	2	50	all	117482.12
she	2025	3	48	all	115879.83
you	1719	4	48	all	98369.10
i	1699	5	49	all	97224.61

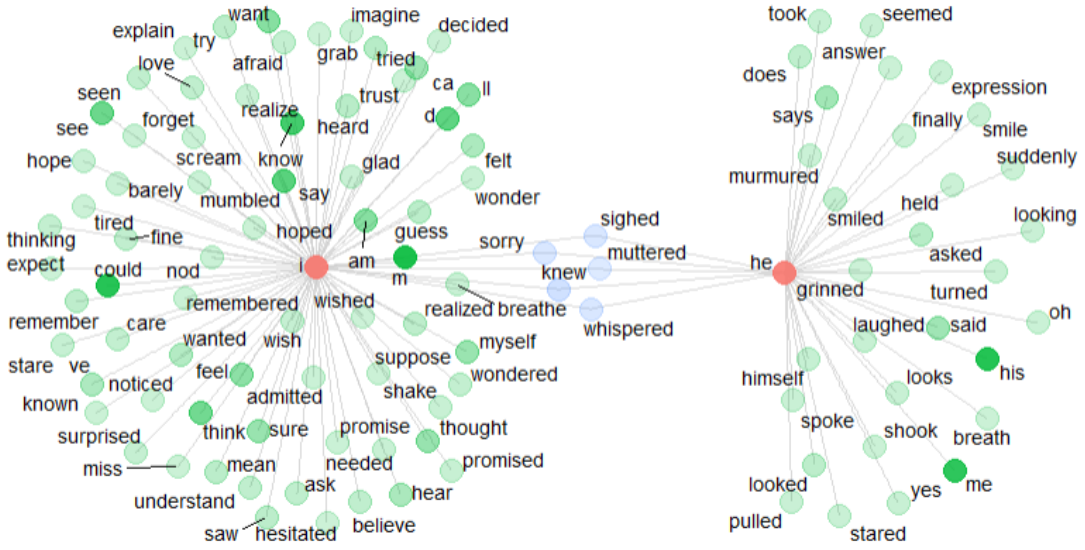
[1] Table 8

From tables 7 and 8, we can see that "i", "he", "it" and "you" are all among the top five most frequent pronouns in each subcorpus. For the young adult pronouns subcorpus, "i" occurs the most frequently with a frequency of 44,109 times and a relative frequency of 334,392.15. For the children's pronoun subcorpus, "he" occurs the most frequently with a frequency of 3,254 and a relative frequency of 186,208.87. We note that "he" and "she" are both in the top five most frequent tokens for the children's subcorpus but only "he" is in the top five most frequent for the young adult. Further investigation into why there is a difference in the frequencies of "he" and "she" between the two subcorpora would be an interesting analysis in a future report.

Next, we analyzed the collocations of "he" and "i" in both subcorpora since these were two of the top five most frequent tokens.

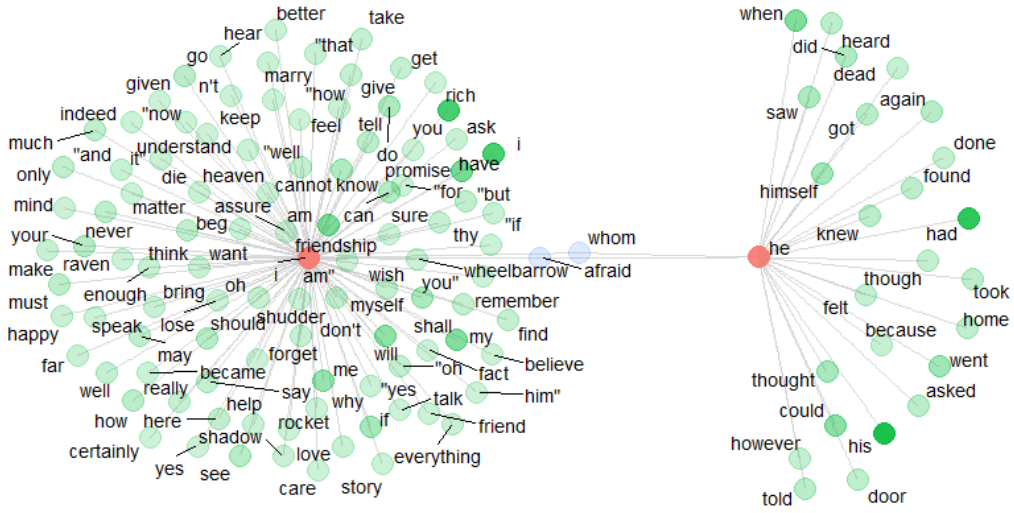
Collocation Network of "he" and "i": Young Adult Body Subcorpus

Figure 8



Collocation Network of "he" and "i": Children's Body Subcorpus

Figure 9



From figures 8 and 9, we see "i" has more words that collocate with it in the children's subcorpus compared to the young adult subcorpus. In the young adult subcorpus, the tokens "m", "know", "realize", "want", "seen". "d,", "could", and "think" collocated the most frequently with "i" and the tokens "hoped", "am", "wished", "wish", "remember", "guess", "m", have the strongest collocation association with "i". In the children's corpus, the tokens "am", "have", "you", "my", and "will" collocate the most frequently with "i" whereas the tokes "can", "am", and "friendship" have the strongest collocation association with "i".

Similar to the collocations for "i", we can also see that more words collocate with "he" in the children's subcorpus than in the young adults subcorpus. In the children's subcorpus, the tokens "himself", "though", "felt", "knew", and "thought" have the strongest collocation associations with "he", whereas, the tokens "when", "had", "himself", "his", and "could" collocate the most frequently with "he". Action verbs and words describing feelings make up a majority of the tokens that collocate with "i" in both subcorpora. In the children's subcorpus, the words that collocate with "he" are more negative (e.g. "dead"), than the words that collocate with "he" in the young adult subcropus which hints at "he" being used when a character is interacting with or describing a love interest.

4.3.3 Looking at Keyness and Effect Sizes

To investigate any statistically significant differences between the two corpora's tokens, we created a keyness table with the young adult corpus as the target and

the children's corpus as the reference. The keyness table includes log likelihood (LL), p-value, and log ratio (LR), as well as deviance proportions, absolute frequency, and relative frequency of a token in the target and reference corpora. The two tables in this section are the same with the only difference being that table 9 is sorted by decreasing log likelihood values, meanwhile, table 10 is ordered by decreasing log ratio values.

Sample Token Keyness Ordered by Decreasing Log Likelihood (LL)

Token	LL	LR	PV	AF_Tar	AF_Ref	Per_10.5_Tar	Per_10.5_Ref	DP_Tar	DP_Ref
i	6315.0485	2.20418	0	44109	1699	5236.2914	1136.31804	0.03686	0.23700
n't	3240.3502	5.58632	0	11098	41	1317.4718	27.42145	0.02848	0.72894
my	1908.0604	1.94249	0	15505	716	1840.6379	478.87211	0.07901	0.24240
s	1812.8413	2.99646	0	9217	205	1094.1735	137.10724	0.05694	0.68621
me	1132.5515	1.71465	0	10799	584	1281.9767	390.58842	0.04865	0.28585
edward	1052.8895	10.15960	0	3222	0	382.4918	0.00000	0.31450	NA
d	752.0046	7.72079	0	2377	2	282.1797	1.33763	0.32371	0.91162
m	698.8408	6.65561	0	2272	4	269.7149	2.67526	0.06100	0.92243
says	693.7386	5.41516	0	2404	10	285.3849	6.68816	0.63982	0.68459
jacob	653.8895	9.47237	0	2001	0	237.5438	0.00000	0.43792	NA

[1] Table 9

Sample Token Keyness Ordered by Decreasing Log Ratio (LR)

	Token	LL	LR	PV	AF_Tar	AF_Ref	Per_10.5_Tar	Per_10.5_Ref	DP_Tar	DP_Ref
6	edward	1052.8895	10.15960	0	3222	0	382.49180	0.00000	0.31450	NA
10	jacob	653.8895	9.47237	0	2001	0	237.54379	0.00000	0.43792	NA
12	bella	451.2850	8.93736	0	1381	0	163.94202	0.00000	0.34895	NA
15	charlie	394.0983	8.74188	0	1206	0	143.16732	0.00000	0.35126	NA
11	re	575.3248	8.32045	0	1801	1	213.80128	0.66882	0.05637	0.98204
22	dauntless	293.4497	8.31644	0	898	0	106.60386	0.00000	0.66135	NA
23	maybe	280.0516	8.24901	0	857	0	101.73665	0.00000	0.09243	NA
24	tobias	274.8231	8.22183	0	841	0	99.83725	0.00000	0.65105	NA
26	carlisle	255.2162	8.11504	0	781	0	92.71449	0.00000	0.34895	NA
34	christina	198.3563	7.75142	0	607	0	72.05851	0.00000	0.65105	NA

[1] Table 10

The token "i" has the highest keyness value, with a log likelihood of 6,315.042. It has an effect size of 2.204, an absolute frequency in the target corpus of 44,109

times, and a dispersion proportion of 0.036 in the target corpus. It is interesting to not that earlier in the report we saw that "i" is also the most frequent token in the young adults corpus. The tokens "n't", "my", "s", and "me" are all among the topy five tokens with the highest keyness values.

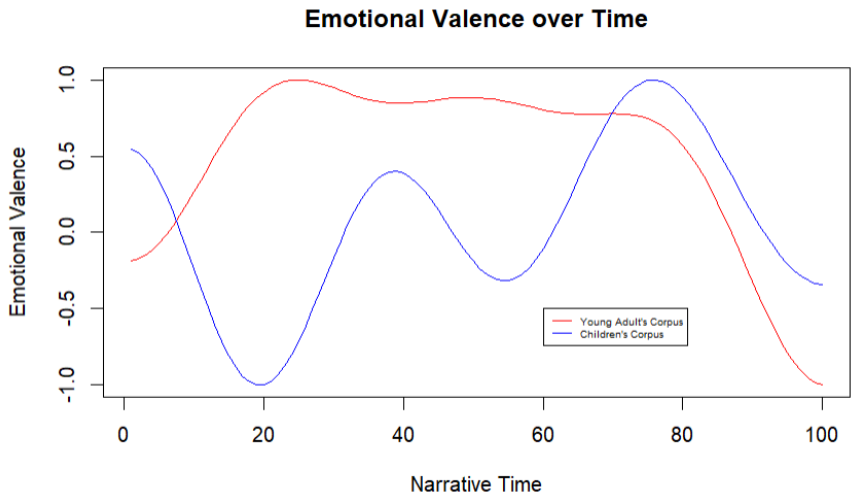
In addition, the token "Edward" has the greatest effect size, with a value of 10.16. It has an absolute frequency in the target corpus of 3,222 and a dispersion proportion of 0.315 in the target corpus. This token most likely occurs at a higher frequency in the target corpus since Edward is the name of the male protagonist in the "Twilight" series which makes up a majority of the young adult corpus.

Since the tokens in the tables all have keyness (LL) values higher than the threshold of 3.84 for significance at $p \leq 0.05$ level and p-values of zero, we can conclude that at least the words in these two tables have significant differences in use between the two corpora. Hence, we can reject the null hypothesis that there is no difference between the frequencies of these words in the two corpora and that all of these words are positive keywords in the young adult corpus.

4.3.4 Looking at Sentiment Analysis

Finally, we conclude our analysis by looking at sentiment analysis to identify if there was any difference in emotional valence between the two copora.

Figure 10



Based on figure 10, we found that the young adult novels stay positive for a longer time than the children's short stories. The short stories start off with negative sentiments and then become increasingly positive. On the other hand, the young adult novels start off positive and then become increasingly negative.

At narrative times of about 8 and 70, the two copora have similar emotional sentiments of neutral (0 emotional valence) and positive (0.75 emotional valence) respectively.

5 Discussion

5.1 Conclusions

We observed a difference in most frequent tokens between the young adult corpus and the children's corpus. Only the tokens "to", "of", and "he" were in common between the two corporas' top five most frequent tokens. Further insepction of body parts and pronouns usage revealed that most of the top used body parts and pronouns for the two corpora were similar rather than different. However, the collocation networks of some of these shared tokens revealed that the contexts in which these words were being used were different. For example, the young adult corpus has more physically descriptive words were collocated with "eyes" than the words that collocated with "eyes" in the children's corpora. Furthermore, the young adult corpus had more postive words collocating with "he" compared to the children's corpus that had negative words relating to death collocating with "he". We postulate that this could be due to the more complex and sometimes romantic nature of young adult books compared to simple short stories made for kids.

Additionally, keyness analysis also found that there is a significant difference in the frequencies of the words used with the young adult corpus as the target and the children's corpus as the reference. Lastly, sentiment analysis found that the young adult corpus shifts from positive emotional valence to negative emotional valence over time while it is the opposite trend for the children's corpus.

In summary, there may be some similar words used between the two corpora, however, the contexts in which they are being used differ between the genres. This is possibly because young adult novels get more dark and complex as the plot line goes on, especially if they are dystopian-themed. Young adult books also portray mature relationships and are often from a first person perspective leading to differences in grammar/language. Children's books on the other hands are short and often non-complex and typically told in third person. Since they typically narrate a life lesson and don't portray as many complex relationships between characters, this would account for the positive emotional valence trend and less usage of descriptive words collocating with frequent words in common with the young adult corpus. Further research into the exact usages and language of these genres would allow us to make more concrete conclusions on the differences that exact between the corpora.

5.2 Limitations

In addition to the results we found, it is important to note some important shortcomings of our analysis incurred to maintain brevity. To begin, if we had done a more extensive analysis we would have liked to enhance how we tokenized our corpora. For example, in the results keyness section we saw that the token "s" has one of the highest keyness values in the young adult corpus, although, "s" is not a meaningful word. We aimed to only keep meaningful and unique words as token but it is possible that how we split and tokenized the corpora resulted in some letters being singled out if they were a part of a contraction.

Our analysis is also limited in its scope since we only analyzed a few features of two extensive genres rather than a wide variety of features to truly hone in on the differences and similarities. Since we also only used two book series for the young adults corpus and fifty short stories for the children's literature corpus (as opposed to a wide variety of young adult series and children's short stories), this further limits the generalizability of our conclusions as these corpora may not be representative of all the books, themes, and motifs in both genres.

Furthermore, we handpicked what words to include in our body parts and pronouns sub-corpora rather than looking at all possible body parts and pronouns that were mentioned in the overall corpora. Hence, we cannot apply the conclusions we draw from this report to words that were not used in our sub-corpora. Lastly, if this report were longer, we would have liked to explore in more depth that the context the body parts and pronouns were used in. Since young adult books are more mature, the words could have been used under different meanings than they are used in children's literature. This is an interesting area for further analysis to dive deeper into what separates the young adult genre from children's literature.

6 References

EDENBD. (2021, July 15). Children Stories Text Corpus. Kaggle. `https://www.kaggle.com/datasets/edenbd/children-stories-text-corpus/`