2016-03-01

Lexical Trends in Young Adult Literature: A Corpus-Based Approach

Kyra McKinzie Nelson
Brigham Young University - Provo

Lexical Trends in Young Adult Literature:

A Corpus-Based Approach

Kyra McKinzie Nelson

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Jesse Egbert, Chair
Mark Davies
Dee Gardner

Department of Linguistics and English Language

Brigham Young University

March 2016

ABSTRACT

Lexical Trends in Young Adult Literature: A Corpus-Based Approach

Kyra McKinzie Nelson
Department of Linguistics and English Language, BYU
Master of Arts

Young Adult (YA) literature is widely read and published, yet few linguistic studies have researched it. With an increasing push to include YA texts in the classroom, it becomes necessary to thoroughly research the linguistic nature of the register. A 1-million-word corpus of YA fiction and non-fiction texts was created. Children's and adult fiction corpora were taken from a subset of the Corpus of Contemporary American English (COCA) database. The study noted differences in use of modals and pronouns among children's, YA, and adult registers. Previous research has suggested that children's literature focus more on spatial relations, while adult literature focuses on temporal relationships. However, the results of this study were unable to verify such relationships. The study also found that YA varied from children's and adult literature in regards to expletives, body part words, and familial relationships. The findings of this study suggest that YA is linguistically distinct from children's and adult. This indicates that future studies should focus more on target audience age. These results could also be applied to L1 reading pedagogy.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: Introduction

Young adult (YA) literature has become one of the most heavily published and widely read subsets of the fiction genre. In 1997, there were 3,000 YA titles published (Brown, 2011). Twelve years later, the number had jumped to 30,000 titles a year (Brown, 2011). The YA title encompasses bestsellers such as the Harry Potter series, *The Hunger Games, Twilight, The Fault in Our Stars, City of Bones,* and *Divergent.* It also includes critically acclaimed novels such as *The Outsiders, The Giver, Speak,* and *The Book Thief.*

Furthermore, within the publishing industry there has been a growing awareness of YA as a genre not only distinct from adult books, but as distinct from books for younger children. Many libraries now shelve YA books in a separate area from other titles. The growing distinction can also be seen in recent changes to the *The New York Times*' bestseller list, which now reports YA and middle grade (books with a target audience of 10-14) as separate categories.

There has also been a greater push to utilize YA books in high school education. *Teaching Young Adult Literature Today* states that "Increasingly, teachers can select well-written young adult titles to effectively engage contemporary students in reading, to get them to care about reading, and as a result, to motivate them and develop more positive attitudes toward reading" (Hayn and Kaplan 42). Educators can now find a number of resources which recommend ways to incorporate YA literature into the classroom. These resources include journals like *The ALAN Review* and books like *From Hinton to Hamlet.*

Despite the popularity of YA literature, the genre has not been given much attention in linguistic studies. If a corpus of YA texts has ever been created, it probably has not been

published or made publicly available. In fact, relatively few linguistic studies on fiction for non-adult audiences exit. As such, very little is known about the differences between adult fiction and YA fiction. Nor do we know much about the differences between YA fiction and fiction for younger readers. Are there lexical and grammatical differences? If so, what sort of differences are there? Linguistically speaking, does YA fiction connect children's and adult fiction? As L1 reading pedagogy continues to push students to read YA texts, it becomes increasingly important to answer these questions and begin to analyze YA from a linguistic standpoint.

With the creation of a corpus of YA literature, we have more opportunity to examine various linguistic features to see how they compare against other subsets of fiction. The purpose of this study is to examine first, whether there are linguistic features (including function and content words) that distinguish YA from literature for other audiences; second, what some of those distinguishing features might be; and third, to determine if the linguistic differences merit identifying YA as a distinct subset of the fiction register.

Definitions

Before continuing, it is important to define exactly what is meant by YA literature. YA literature is literature intended primarily for a 14 to 18-year-old audience, primarily intended in the sense that while adults can (and often do) read YA literature, they are not the target audience. In fiction, this almost always means that the main character falls into the 14-18 year age range. Additionally, YA is often distinguished from books for younger audiences by its inclusion of more mature themes. In this paper, the term juvenile fiction is used as a blanket term to encompass literature for children, preteens, and teens.

While the majority of books are easy to classify as either YA or not, there are some titles which evade easy classification. For instance, books like *The Curious Incident of the Dog in the Night-time, Ender's Game,* and *The Catcher in the Rye* all feature teenaged protagonists, but were not necessarily intended for a teen audience when published, which causes confusion as to how they should be categorized. Conversely, Rainbow Rowell's *Fangirl* features a protagonist of college age, but was marketed towards teens.

Despite these exceptions, the majority of books are easily classified. The books included in this corpus were all classified as YA by numerous users on the Goodreads website. This classification system will be discussed further in the methods chapter.

CHAPTER TWO: Literature Review

Corpus based studies on register variation

According to Biber, Conrad, & Reppen (1998) corpus methodology is an empirical research method that depends on both quantitative and qualitative analysis. It also "utilizes a large and principled collection of natural texts, known as a 'corpus' as the basis for analysis" (pg. 4). Many previous studies have used corpus methodology to investigate issues in register variation. Register describes "varieties defined by their situational characteristics" (Biber, Conrad, & Reppen 1998). Kennedy (1998) documents the importance of register variation in *An Introduction to Corpus Lingusitics.* Biber (2012) has noted that collocates of high frequency words behave differently based on register and that collocation was only one of many features that varied between registers.

Corpus methods have been used to document register variation for a wide assortment of registers, especially in recent years as computers have become better equipped to analyze large volumes of text. Parodi (2013) found that within the textbook register, there were discourse differences between different disciplines. Another study by Grabowski examined differences across pharmaceutical texts (2013). Quaglio has used corpora to compare dialogue in the sitcom *Friends* to natural conversation (2009). Register variation has also been documented in languages other than English, including Chinese (Zhang 2012), Brazillian Portuguese (Sardinha, Kauffmann, and Acunzo 2014), Korean (Biber and Finegan 1994), Somali (Biber and Hared 1992), and Gaelic (Lamb 2008). This is only a small sampling of studies on register variation that have been performed using corpus methods. Really, any registers for which a corpus can be created are open for investigation using this methodology.

Adult Fiction

Despite the extensive research that has been done on register variation, fiction has not attracted as many corpus-based studies as some other registers. However, there have been some notable studies. This may be due in part to the fact that fiction is highly protected under copyright, which can make it more difficult to obtain in a searchable format. Creation of a fiction corpus generally requires either hours of scanning books or pirating digital copies.

Even with these limitations, there are studies that have focused on a specific subset of fiction. For instance, Siepmann (2015) focused on post-war fiction, while Mahlberg (2010) looked at nineteenth-century fiction. Corpus methods have also been used to examine the texts of individual fiction authors. Mahlberg (2012) used corpora to analyze the writings of Charles Dickens. Her use of corpus stylistics has allowed a quantitative measure of some of the stylistic features noted by other Dickens scholars. Dossena (2012) has used a corpus-based approach to studying the fiction of Robert Louis Stevenson. Some research has even used corpus methods to look at a single text by an author, such as one study from Fischer-Starcke (2009) which analyzes Jane Austen's *Pride and Prejudice.* It may be noted that most of these texts are old enough to evade copyright issues, so the texts are readily available in a searchable format. Studies focused specifically on modern fiction are more difficult to find.

Young Adult Literature

Because the emergence of YA literature as a recognized genre is a fairly recent phenomenon, relatively few studies have specifically targeted YA. While there is certainly room for more YA research across disciplines, there has been more research on YA produced by literature and education disciplines than linguistics.

The Assembly on Literature for Young Adults (ALAN) was founded in 1974. The Assembly publishes *The ALAN Review*, a triannual journal of peer-reviewed articles focusing on different aspects of YA literature. The articles published in ALAN and similar publications typically note a trend and give several examples of books following the trend. For instance, Cox noted that first person present tense narrative has become increasingly common in YA literature, and cites Andrew Smith's *The Marbury Lens* and Daria Snardowsky's *Anatomy of a Boyfriend* as examples (Cox, 2013). Cox also concludes that use of first person present narrative adds a sense of urgency and immediacy to the story. While some articles such as Cox's discuss more linguistic features like first person narrative, most tend to focus more heavily on themes and content in YA books. For instance, Brown and Crowe (2013) coauthored an article discussing sports in YA literature. In another article, Durand explores post-colonial YA literature (Durand, 2013). These studies often make interesting qualitative observations about the nature of the YA register, but lack a quantitative element. Furthermore, they generally focus on what is occurring within YA without comparing trends against those found in books for older or younger readers.

These studies can be useful for finding features we would like to measure quantitatively. Of course, qualitative studies can be valuable and certainly have their place. Qualitative studies are particularly useful in analyzing single texts or small groups of text. Qualitative observations may also be used to provide assumptions which can then be tested qualitatively. Quantitative data allows us to test our assumptions to see if they are true. Furthermore, quantitative studies allow us to look at whether a feature is really being used broadly across a discipline, rather than in a handful of works which have been used as examples of the feature.

L1 reading Pedagogy

Current pedagogy in L1 reading and vocabulary instruction emphasizes student-motivated reading programs. Exposure to new words is critical for vocabulary gains, and research has shown that after third grade, students are exposed to the majority of new words through reading (Gardner, 2004). Because of the critical role reading plays in vocabulary acquisition, current pedagogical approaches favor wide-reading (Nagy, Herman, and Anderson 1985). The theory behind this approach is that the more a student reads, the more words they will be exposed to and be able to acquire. Working hand-in-hand with this method of instruction is a focus on student-motivated reading. Current pedagogical practices prioritize helping students learn to love reading. By fostering a love of books, teachers hope that students will naturally read more, leading to the vocabulary gains anticipated by the wide-reading approach.

The past several decades have also seen an expansion in the publishing of juvenile literature. More books for young readers are being published and sold each year. Many educators are responding by pushing for more use of juvenile literature in the classroom. (Herz & Gallo, 1996) Also of note, publishing has seen a growing awareness of age gradation. What used to be a blanket audience of "juvenile literature" has evolved into more fine-tuned categories such as board books, picture books, early readers (target age 4-8), chapter books (target audience 6-9), middle grade books (target age 10-13), and young adult books (target age 14-18). These categories may be even further divided, for instance distinguishing between lower YA (ages 14-16) and upper YA (ages 17-18). Yet despite this expansion and the increased push for students to read juvenile literature, many questions remain regarding exactly what vocabulary students are being exposed to in these books. Very few linguistic studies have been done with the aim of gaining a quantitative understanding of the language of juvenile literature, and those that have been done mostly ignore distinctions between the different target audiences.

Support for wide reading approaches date as far back as St. Augustine (Nagy, Herman, and Anderson, 1985). Proponents of this approach rally around the Incidental Acquisition Hypothesis which states that most vocabulary gains are made through natural encounters with the language. Grade school students learn a large amount of vocabulary and explicit instruction alone cannot account for this vocabulary growth. So it is assumed that most vocabulary gains are made through repeated exposure to words. In addition to being exposed to words, learners must have the skills necessary to determine the word's meaning from context.

Despite the vast support of wide reading, some concerns remain. For instance, if vocabulary gains are based on exposure, how many instances of exposure are necessary for a reader to learn a word? How many times must the word be encountered before it is learned? Also, how well are words being learned?

The number of necessary exposures is likely influenced by the helpfulness of the contexts it is found in. If no direct vocabulary instruction is received, the reader is left to learn or acquire the meaning of a word from surrounding clues. There are a number of different clues which may be used. Dubin and Olshtain (1993) detail a number of factors which may contribute to a reader's ability to derive meaning from context. For instance, extratextual information, or the reader's general knowledge extending beyond the text, may be a factor in guessing word meaning. Semantic knowledge, both on the sentence and paragraph level and on the larger level of discourse, may also influence acquisition through context. Furthermore, thematic understanding of the text can aid readers. In other words, how well do readers understand the rest of the content? Finally, syntactic clues can help readers understand meaning.

Furthermore, morphological clues may be found within the word itself. Studies suggest that students who are morphologically aware are better able to decode word meanings, increasing

the likelihood of their learning new words encountered in unstructured reading (Pacheco & Goodwin 2003). A study of middle school readers also showed that students were more likely to make morphological connections after being explicitly taught morphological strategies (Pacheco & Goodwin 2013). The study also found that morphological strategies could be used to deepen knowledge of words students already know.

Illustrations can also be a useful source of context, particularly in children's books. Use of eye-movement software has better enabled researchers to study the connection between illustrations and vocabulary. One notable study examined the eye-movements of four-year-olds who were read an illustrated story multiple times (Evans & Saint-aubin, 2013). The study showed that students fixated on the portion of the illustration that was being mentioned in what was being read aloud. They also showed that after multiple exposures, students began looking at the text itself more, although they were all pre-readers. Students were given pre and post vocabulary tests, which indicated that vocabulary gains were made after multiple readings.

While the research does indicate that children can make significant reading gains through input alone, studies have also shown that explicit instruction can be useful for helping students make even wider vocabulary gains. Gonzalez et al. (2014) studied 100 children taught by 13 teachers over the course of 18 weeks to analyze the relationship between teacher talk and vocabulary gains. The study found that teacher interaction with students before, during, and after reading did affect student vocabulary gains. Teachers who spent more time on extratextual talk were able to see greater vocabulary gains in their students.

With all this in mind, we can now turn our attention to understanding what types of lexical items students may be exposed to in the texts they read.

Variation within Juvenile Fiction

Linguistic studies, particularly vocabulary studies, often focus on differences between different registers. We expect variation in genre to manifest itself lexically, and this holds true for juvenile literature. Gardner (2008) has noted differences differences between expository and narrative children's texts. Research indicates that vocabulary between expository and narrative texts differs even when the texts are clustered around a common theme (Gardner, 2008). Gardner also found that expository texts recycled more specialized vocabulary than thematically similar narrative texts.

This distinction becomes important when addressing the pedagogical approaches surrounding wide reading. While students may make vocabulary gains by widely reading self-selected narrative texts, they will not be exposed to the same types of vocabulary they would be exposed to with expository texts. Additionally, the words would be repeated less, decreasing chances for vocabulary gains. This becomes problematic as the lexical items that are more specific to expository texts are the types of specialized vocabulary essential to academic success.

Furthermore, research shows that children's literature differs linguistically from adult fiction. In order to better examine what words appeared most frequently in children's books, Stuart et al. (2003) created a database of 685 books for children ages 5 to 7. They then were able to create frequency lists. They noted that previous lists were inadequate because they either consisted of adult language or American language (as opposed to British English or other world Englishes). While the wordlists from American children's books may have been more accurate, it was also problematic in that it was created in 1971, making it rather dated. Stuart's team noted that there were a large number of nonwords, particularly interjections. They also found that the most frequent words were function words rather than content words, which is not a finding unique to their study. When looking at gendered pronouns, they found that male pronouns were

significantly more prevalent than female pronouns. While the study has a number of merits, it could be improved upon by taking care to indicate how the features they found contrast with features in adult literature.

More recently, researchers at Oxford created a 30 million token corpus of texts for children ages 5-14, including both fiction and non-fiction (Wild, Kilgarriff, & Tugwell, 2013). They performed a keyword analysis against a corpus of adult texts and found a number of differences. Some of these differences were fairly predictable. For instance, they found that children's literature contained lexical items that correlated strongly with the physical, concrete words, while adult fiction tended to have more abstract terms. For instance, body parts, buildings, tools, and landscape related vocabulary correlated more strongly with children's literature. Words relating to religion, politics, business, and education correlated more strongly with adult fiction. Furthermore, keywords revealed that children's literature focuses more on relationships between siblings and parents while adult fiction focuses more on relationships between romantic partners and children. Predictably, the study found that, on average, words in children's books were shorter—an average length of 4.7 characters compared to adults 6.2 characters. As in Gardner's study (2008), the keyword lists showed differences between the vocabulary in expository and narrative texts for children.

Several less predictable differences were noted as well. Children's literature focused more on spatial relationships (demonstrated by the keyness of words like *bottom, hole, shape, edge,* and *gap)* while adult adult fiction focused more strongly on temporal relationship (demonstrated by words like *late*, dates and ordinals which appear on the adult side). Modals and auxiliaries also seemed to be more common in children's literature than adult literature, though the authors do not give any explanation for why this may be.

CHAPTER THREE: Methods

Corpus construction

Corpus research has become one of the most popular methodologies in linguistics. Yet, it is important to remember that corpus results are a product of the corpus they come from. As Douglas Biber (1993) points out:

> The use of computer-based corpora provides a solid empirical foundation for general purpose language tools and descriptions, and enables analyses of a scope not otherwise possible. However, a corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning a language as a whole.

As such, it is important to pay attention to corpus construction to prevent, to the extent that it is possible, skewed or misleading data.

In this study, young adult literature is broadly defined as texts with an intended audience of readers ages 14-18. In fiction, the primary protagonist will also generally fall in this age range. This definition of YA literature conforms to what is widely accepted in the publishing industry. This guideline is also typically used in determining how to categorize books for awards, where to shelve them in book stores and libraries, and which grade levels they should be taught in. The books in the corpus are generally more modern (published in the last ten years) however some older books were also included, dating as far back as 1967 (S.E. Hinton's *The Outsiders).* This is important to note as language varies not only across registers, but across time as well.

The Young Adult Corpus (YAC) contains 1,005,147 words pulled from 67 YA books. Of these, 52 titles are fiction and account for 773,771 words in the corpus. Books were

selected from a list of popular teen fiction on Goodreads. The Goodreads list is derived from votes from a large reading community, and reflects which books are most popular. Users on the Goodreads site can add books to the list as well as vote on books already on the list. Books with the most votes appear at the top of the list. Based on this system, books that are more widely-read are likely to float to the top of the list and are more likely to be included in the corpus.

Every third book on the list was chosen for inclusion in the YAC. However, only one book by any given author was included. Occasionally, a book on the list would be unavailable from the library, thus a small number of these books could not be included in the corpus.

Once the books were selected, 60 pages from each text were scanned and converted to text using Adobe Pro. The converted texts were then saved as .txt files. The titles of the books were used as filenames, which helped easily identify texts in this study, but might prove confusing if the corpus were expanded to include more books.

The 60 pages consisted of twenty pages from the beginning of the book, twenty from the middle, and twenty from the end. In a few cases, additional pages were scanned if the book contained illustrations or other graphics. This was done to ensure that the overall number of words drawn from each text was fairly consistent. Although time constraints made it impossible to scan full books, my hope was that by scanning from the beginning, middle, and end, I would be able to get the most accurate representation of the text.

On average, 15,000 words were taken from each book. Most were near this average, while there were some outliers (20,000 at the high end and 7,000 at the low end). This distribution suggests that most individual books account for only 1.5% of the words in the

corpus and no individual book accounts for more than 2% of the corpus. This should mitigate the ability of a single text to skew results.

After the scanned files were converted to readable text, they were checked to make sure they were generally correct. Some editing was performed to clean up OCR errors. Many of the errors occurred in high frequency words and were predictable within texts. For instance, a certain novel may use a font where the OCR software would consistently confuse *if* as *lf*. In such cases, the find and replace feature was used to quickly identify and fix errors. The spellcheck feature was also used to identify a number of errors, many of which were caused by the OCR software being unable to perceive a space between two words. These were also easily fixed. There are most likely OCR errors remaining in the corpus, particularly for words that are spelled incorrectly as other real words (such as 'am' being read as 'an') because the spellcheck would not be able to pick these up. However, the majority of words were read correctly. The text still makes sense and the existing errors do not make it difficult to read, which suggests that it is suitable for research.

Beyond being sorted into fiction and nonfiction, no attempt was made to control for genre (genre here being used in regards to different subsets of fiction such as science fiction, mystery, historical fiction, etc.). While studies of adult fiction have shown that genre can have significant impact on vocabulary, and while an analysis on differences between genre in YA would be interesting, it was beyond the scope of this project to examine such features. Despite not controlling for genre in the sampling process, all major genres (fantasy, science fiction, contemporary, historical, biographical, and informational) are represented, along with a variety of subgenres.

Comparative corpora

In order to determine if YA differs from adult and children's literature, it was necessary to have corpora to compare the YA corpus against. This was achieved using two sub corpora drawn from the fiction portion of the Corpus of Contemporary American English (COCA). A million words of children's fiction were used to create one of the corpora, and a million words of adult fiction were used to create the second corpus.

It is important to note one major difference between the adult and children's corpora used and the YA corpus created for this study is that the former use only material drawn from first chapters. This may have a slight affect on some of the searches, and should be taken into consideration when analyzing results.

Keyword analysis was performed to determine which words were particularly salient in YA fiction, as opposed to children's fiction and adult fiction. Use of keywords has been used in previous research, such as the Oxford study (Wild, Kilgarriff, & Tugwell, 2013). There are two steps to creating a keyword list. First, every word in the corpus is compared against a reference corpus to find out how common it is in the main corpus compared to the reference corpus. Then the ratios of these words are ranked. Doing this allows us to determine which words are used substantially more in the main corpus as opposed to the reference corpus. For this study, keywords were found using AntConc's Keyword list features. The adult and children's corpora were used as reference corpora to create two separate keyword lists. After loading in the reference corpora, keywords were generated using log likelihood.

Features examined

Two methods were used for determining which features to examine with the corpus. The majority of features examined in this study were also analyzed in the Oxford study (Wild, Kilgarriff, & Tugwell, 2013). This was done to see if the results from that study could be replicated using different children's and adult corpora. Also, I hoped to build on the original study by including YA fiction. The Oxford study noted differences in use of modals, animal words, temporal words, and spatial words. They also noted that the children's literature and adult literature focused on different types of familial relationships. All of these features were examined using the corpora created for this study.

In addition to examining features studied by the Oxford group, I also chose some features to examine based on results from a keyword analysis run in AntConc. The keyword lists themselves were messy—largely due to the fact that contractions are tagged differently in COCA than they are in the YA corpus, and this caused any frequent contractions to appear higher on the keyword list than they should have been—however, they did provide several interesting words which prompted further investigation. While the keyword lists themselves were not very useful, they did help me decide to look more deeply into use of pronouns and body part words, both of which yielded interesting results.

CHAPTER FOUR: Results and Discussion

After creating the corpus, a number of features were examined to see how they compared across registers. Some of the features examined were from the Oxford study, to verify results between children and adult, as well as to see how YA compared. Other features were examined after a keyword analysis suggested there might be some interesting phenomenon occurring with the feature.

Modals

One of the findings from the Oxford study suggested that modals are more common in children's literature than adults. This claim was examined with the corpora used in this study. Additionally, modals in YA were compared against modals in children's and adult books.

The findings presented in Figure 1 confirm that overall counts for all modals were highest in children's and lowest in adult, with YA fiction falling between the two.
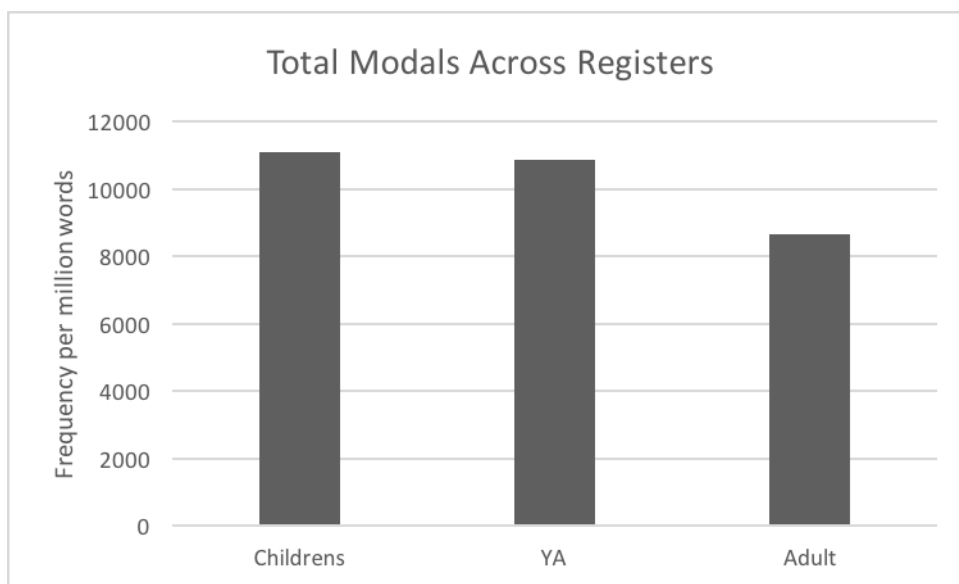
Figure 1: Modals total

However, a closer examination of Figure 2 shows that individual modals varied substantially in their use across registers.
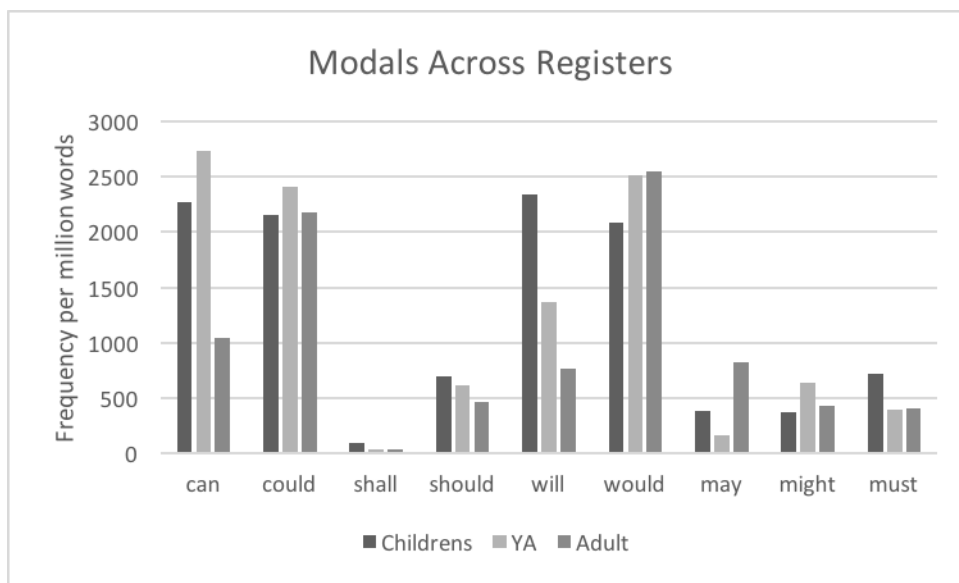
Figure 2: Modals break down

To draw some conclusions about why these differences exist, it is important to consider the functions of modals. Modals are meant to indicate permission, volition, possibility, and ability. Certain modals are more likely to fulfill certain functions than others.

Can and *could* appear to be most common in YA. These most often appear to be used to denote ability, though they can also be used to request or grant permission. This suggests that YA and children's literature seems to focus on character ability, which may reflect the way in which children and teens approach their emerging identities.

Can and *could* in YA:

1. *Alright, I can keep up.*
[Cress]

2. *I can do this by myself.*
[Flipped]

3. *I thought you only could control the weather.*
[I am Number Four]

4. *I relaxed. We could take whatever was coming now.*
[The Outsiders]

Shall and *should* are most commonly used when giving advice or direction. It makes sense that these would be most common in children's literature, which tends to be more heavily centered on teaching morals than literature for teens and adults. These, in addition to *must*, might be lower in YA, since teens are generally less interested in being told what to do.

Shall and *should* in children's literature:

1. *You shall sell that worthless property.*
2. *You should have asked me first.*
3. *Maybe you should think twice.*

Body Parts

One trend that seems notable is that words for various body parts appear more in YA fiction than either adult or children's fiction. Adult fiction was also consistently higher than children's fiction (with the exception of the word *arm*) but never as high as young adult. This seems to suggest that YA is very physical in nature. In particular, *hand, face, and lips* were higher than adult in fiction, which may be due in part to the central role romance plays in YA fiction.

A look at the words in context reveals that often these words are used in order to advance a romantic subplot.

1. *His eyes were as intense-and as gold -as she remembered.*
[Dangerous Creatures]

2. *He bowed low to kiss Raisa's hand.*

[The Grey Wolf Throne]

3. *He smoothed her hair off her face.*
[Eleanor and Park]

4. *He takes a shaky breath and pulls me close. Kisses the top of my head.*
[Shatter Me]

5. *I can just make out the lines of him, and, of course, feel the warmth from his skin.*
[Before I Fall]

6. *When she reached her fiancé, he took her arm and led her back to the landau.*
[The Luxe]

However, romance subplots (or even main plots) are not enough to account for all the body words in YA, as there are a number of examples of all of the words being used in non-romantic contexts. Quite frequently, body words appear in beats, the actions used in place of a dialogue tag to indicate speaker. There are also many references to characters being injured which use body words.

1. *I picked it up even as it burned hot and its edges sliced my hand.*
[Hex Hall]

2. *I close my eyes, willing it all to go away.*
[A Great and Terrible Beauty]

3. *No other words formed on my lips.*
[Blood Promise]

4. *She has spotted Adam through all the other invaders and her face has gone pink with anger.*
[If I Stay]

5. *I pull the tissue away from my face. Blood drips.*
[Need]

6. *Perry shook his head in disbelief.*
[Under the Never Sky]

7. *Here are my bad traits: a too-long nose, skin that gets blotchy when I'm nervous, a flat butt.*
[Before I Fall]

8. *I feel calm as I undo the braid in my hair and comb it again.*
[Insurgent]

9. *My broken arm jostles. My teeth clench.*
[Need]

The word *blood* may also be of interest because it is rarely, if ever, used in a romantic context. However, it is considerably more frequent in YA. Even taking some skewing into account (one book is titled *Blood Promise*, and the title of the book was sometimes included on the scanned pages), the word appears much more frequently in YA, suggesting that even beyond romantic plots, YA is more physical in nature.

There are also a number of cases where body words are used to describe aspects of appearance that the characters are not fond of, for instance in example sentence 7. This coincides with teens' growing awareness of their bodies and the insecurities that often accompany that awareness.
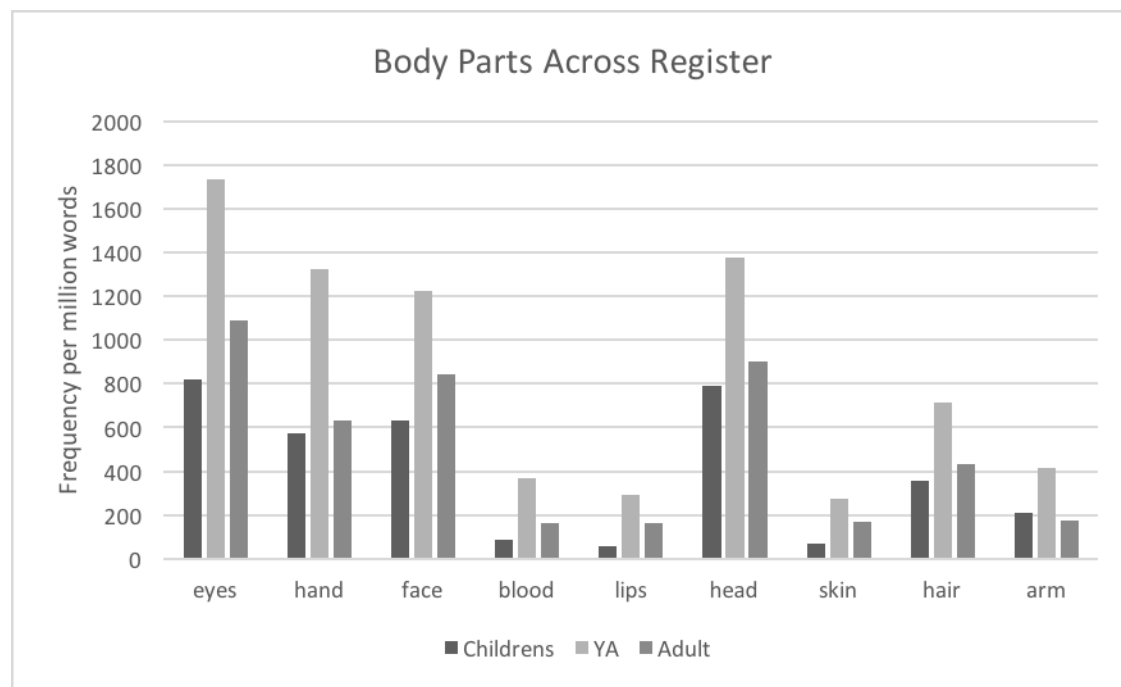
Figure 3: Body Parts

Pronouns

An examination of the differences in pronoun use across registers reveal some interesting insights about register variation.

One of the most stark differences lies in the high use of *I* in YA literature. This suggests that YA literature makes greater use of first person narrative styles. This gives quantifiable evidence to support the claims that YA seems to embrace first person (cox, 2013).

Difference in use of feminine and masculine pronouns may also be of interest. For all registers, feminine pronouns were less frequent than masculine pronouns. The difference is most pronounced in children's literature. On the surface, it may be surprising that YA has more masculine pronouns than feminine, when YA is often considered to be more geared toward female audiences. One possible explanation for this is that many YA books feature a female protagonist and use first person narration. In these cases we would expect that first person, gender neutral pronouns are being used to refer to the main (female) character. In these cases, supporting characters (including the love interest) are more likely to be male and referred to using masculine pronouns. This does however substantiate claims by those who propose that YA should depict more relationships—including sisters, friends, and love interests—between girls.

There are also some interesting patterns in regards to children's literature. Just as use of *I* in YA suggests more first person narration preferences, high use of *you* in children's may denote a preference for second person narration.

Children's literature also has the highest use of the pronoun *we.* This may suggest that children's literature puts a stronger emphasis on themes such as teamwork and working together than YA or adult literature.
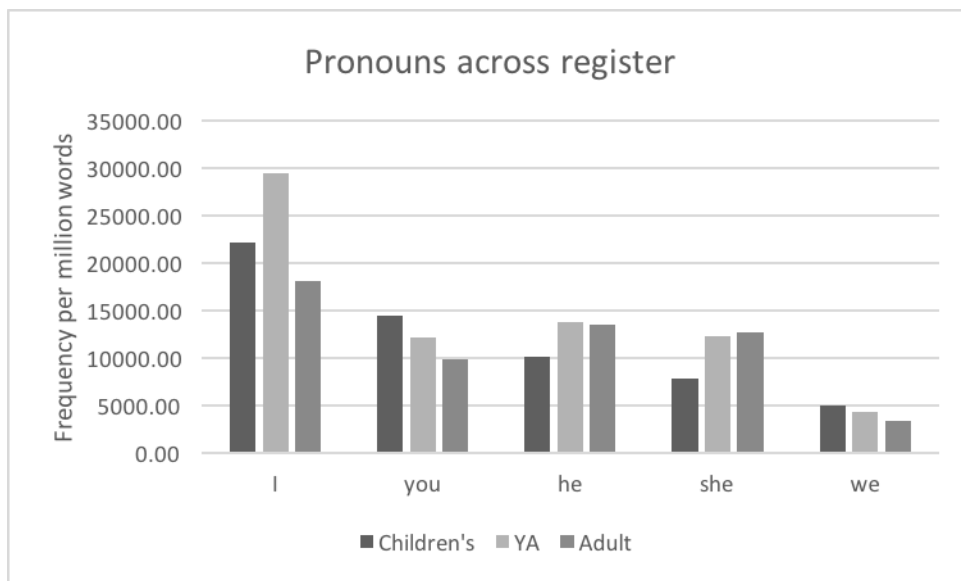
Figure 4: Pronouns

Time words

One of the findings of the Oxford study suggested that words related to time were more common in adult fiction than children's fiction. In this study, several time words were studied to see how they compared across registers.

The majority of time words looked at were most common in adult, though the patterns did not seem as clear as the Oxford study seemed to suggest. Several words were higher in children's literature or were very close to their counts in adult literature.

The word *now* was most common in YA, followed by children's. *Soon* was also more common in children's with similar counts in YA and adult. These findings may suggest that

books for teens and children may focus more on what is immediately happening. In YA, the

high use of *now* may indicate the common use of present tense narration.

Now used with present tense narration:

1. *Her eyes are open now.*
[Matched]

2. *Emma is done with French now and unpacking her violin.*
[Wintergirls]

The word *never* was more common in adult, which may indicate that children's

books typically have a more positive tone than adult books.
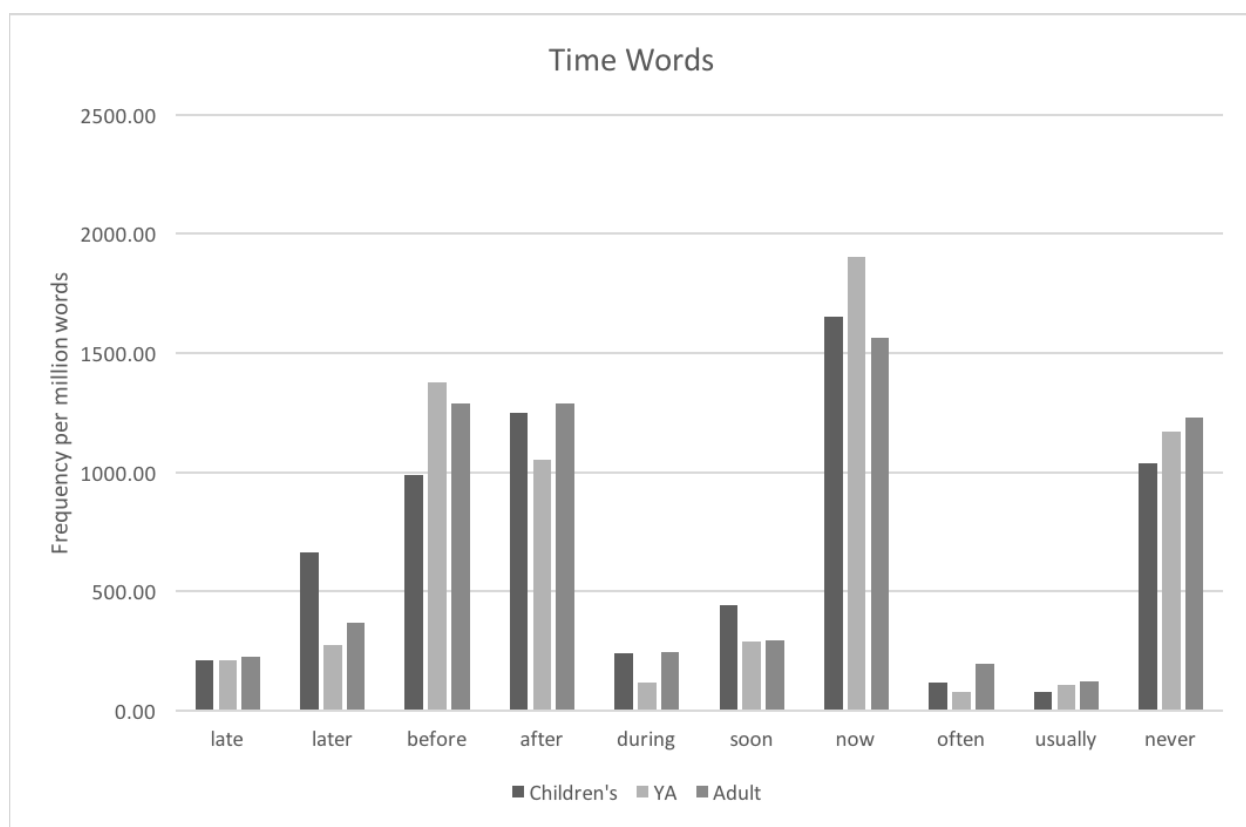
Figure 5: Time words

Expletives

One of the ways in which YA literature differs from children's literature is the presence of more mature content. This includes profanity. While profanity is generally unacceptable in children's books, it is used in YA books. The findings from this study show that expletives are used in YA fiction; however, they are not as common as they are in adult literature. Profanity was nonexistent in the children's texts that were examined. While there were a few hits for the word *hell*, all instances were literal.
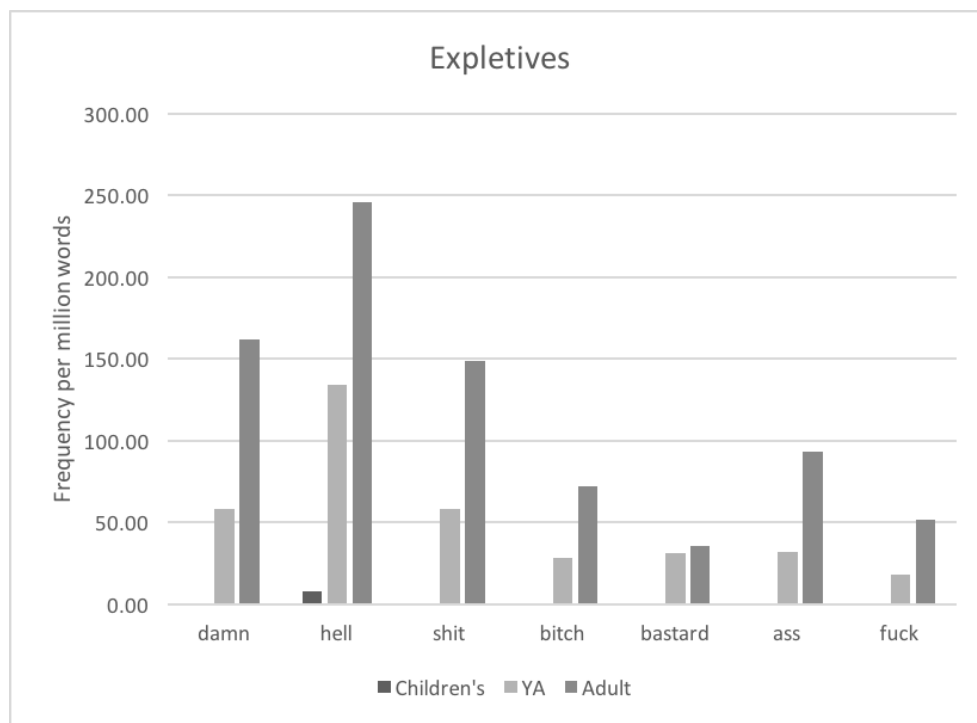
Figure 6: Expletives

Animals

The Oxford study found that animal words were generally more common in children's than adult. This seems true based on my data as well. YA was, on the norm, considerably lower than children's or adult, perhaps because it is actively trying to seem less childish.
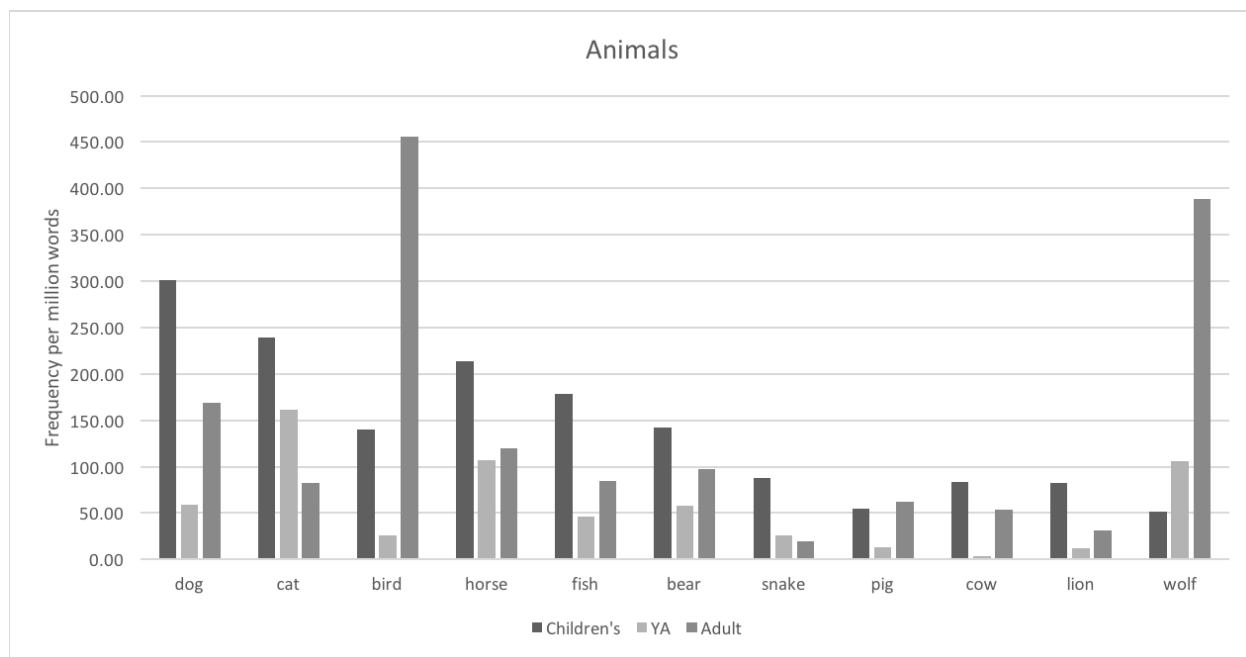
Figure 7: Animals

Upon examining the data, it appears that there is some heavy skewing for *bird, pig,* and *wolf.* The adult texts contained books with characters named Whippy Bird, Pig Face, and Wolf. There was also some skewing for *wolf* in YA, where one character was named Wolf and another book had a title with *wolf* in it. With these skewings in mind, it becomes even more clear that children's literature has much more reference to animal words. It also seems that adult literature is more likely to use animal words in a metaphorical sense or with idioms.

Furthermore, children's literature has a much higher likelihood of treating animals of characters or having characters who actually are animals. In adult and YA, they are more likely to be relegated to the realm of pets. Often the animals in children's literature will talk, while talking animals are difficult to find in adult or YA literature.

Metaphoric examples from YA:
1. *Why am I hopping around like some trained dog trying to please people I hate?*
[The Hunger Games]

2. *His reflexes were as sharp as a cat's.*
[The Icebound Land]

3. *You're the most disgusting piece of pig lard I've ever seen.*
[The Infinite Sea]

4. *He slipped the horn into his breast pocket, then rose from his lion-haunched crouch and turned.*
[Daughter of Smoke and Bone]

Example sentences showing the personified nature of animals in children's books:
1. *Look at Tiger and Dog go!*
2. *"I'm not a street cat. I'm a house cat."*
3. *The white snake warned Lien that strangers were approaching the forest.*
4. *I've never driven a cow to a party.*
5. *If he didn't, Lion would tease him all day.*

Spatial words

One of the key findings from the Oxford study suggested that while adult literature focused more on temporal relationships, children's literature focused more on spatial relationships. They suggested that this indicated that children's literature was more focused on the physical world than adult literature. The study listed several key words from their children's corpus which supported their claim. Those words were examined in this study.

Results from this study differed with the findings of the Oxford study. In fact, these results seem to contradict the findings from the Oxford study as only one of the words examined was greater in adult fiction than children's fiction.
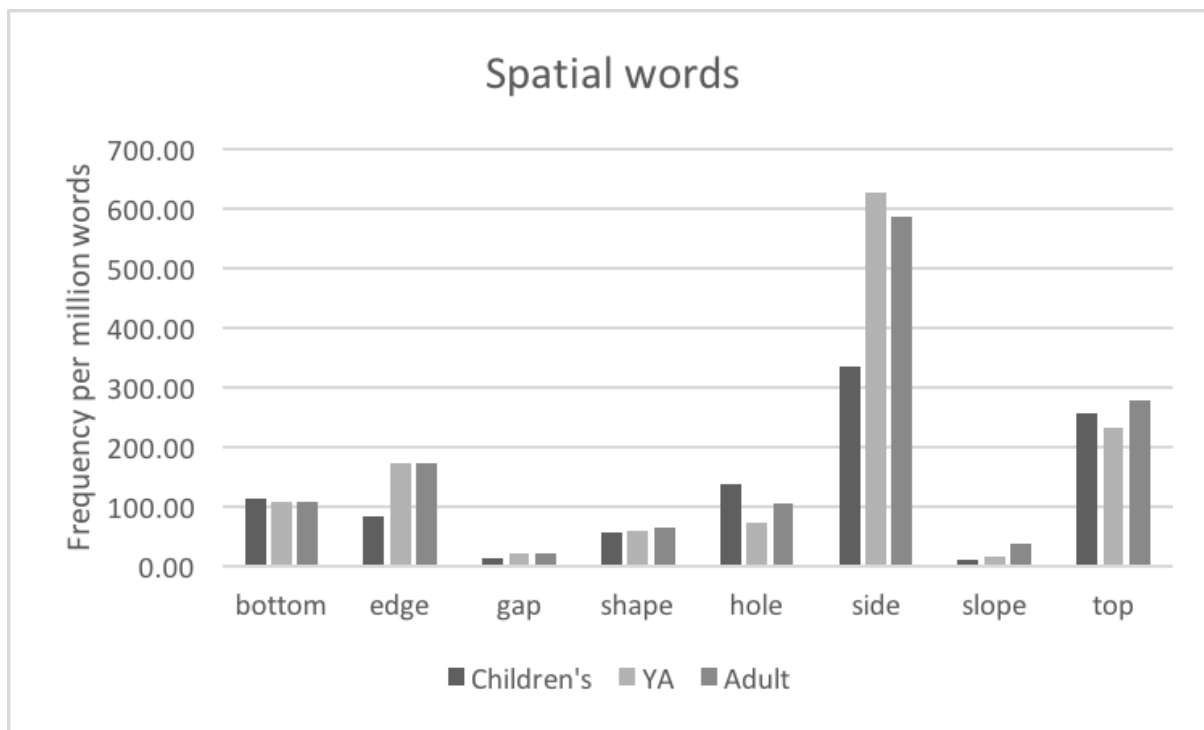
Figure 8: Spatial words

Parental relationships

Overall, parents are mentioned most commonly in children's fiction, followed by YA with adult mentioning parents the least. However, words used to talk about parents varied considerably depending on the register.

For all registers, mothers were mentioned more frequently than fathers. Words for parents decreased consistently as target audience increased, suggesting that as characters grow older their focus shifts to other relationships. When looking at the breakdowns of specific words; however, it appears that more formal forms such as *mother* and *father* are most common in adult. They were also more common in YA than in children's. However, all of the less formal forms (such as *mom, mommy, mama, dad, daddy, papa)* were more common in children's literature than any other register. Forms such as *mama* and *mommy*

were especially uncommon in YA. This is likely because it sounds too childish for a teen to use. These forms, while still infrequent, may be slightly higher in adult because adult characters may have children who refer to them to them with those terms, something teens are not likely to have.

These relationships make sense. Children typically live with their parents and rely heavily on them. Teens are more independent, however they often still live at home and rely somewhat on their parents. Adults, however, usually do not live with their parents anymore. Also, many of the hits for parental words in adult fiction are actually cases where adults are talking to children about their parents.
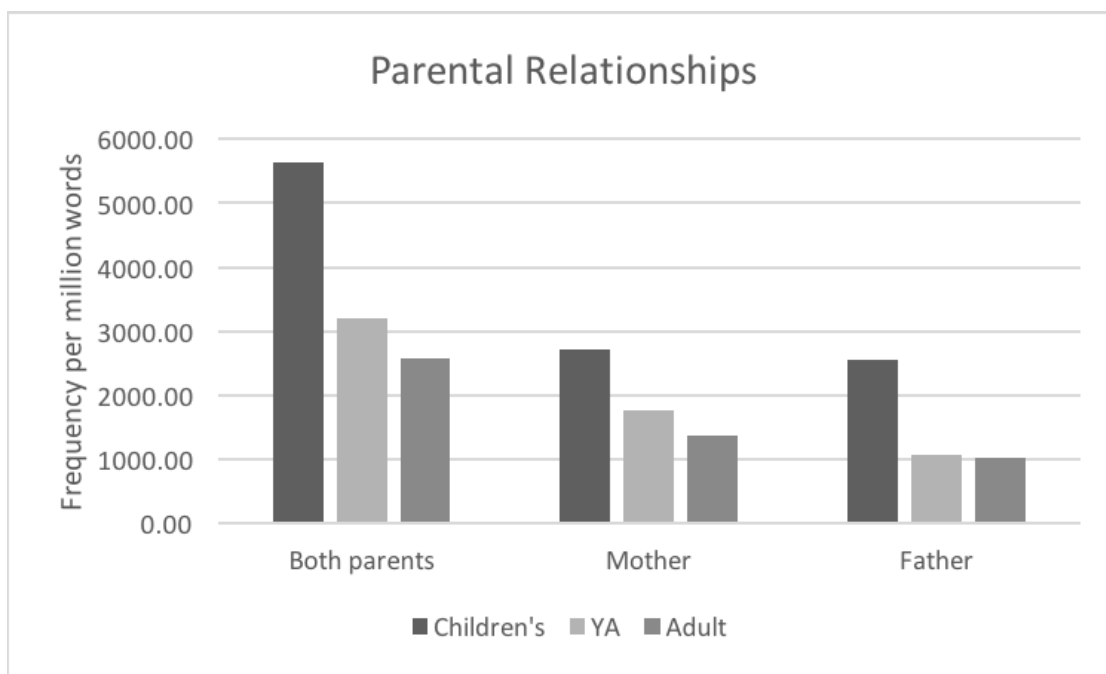
Figure 9: Parental relationships
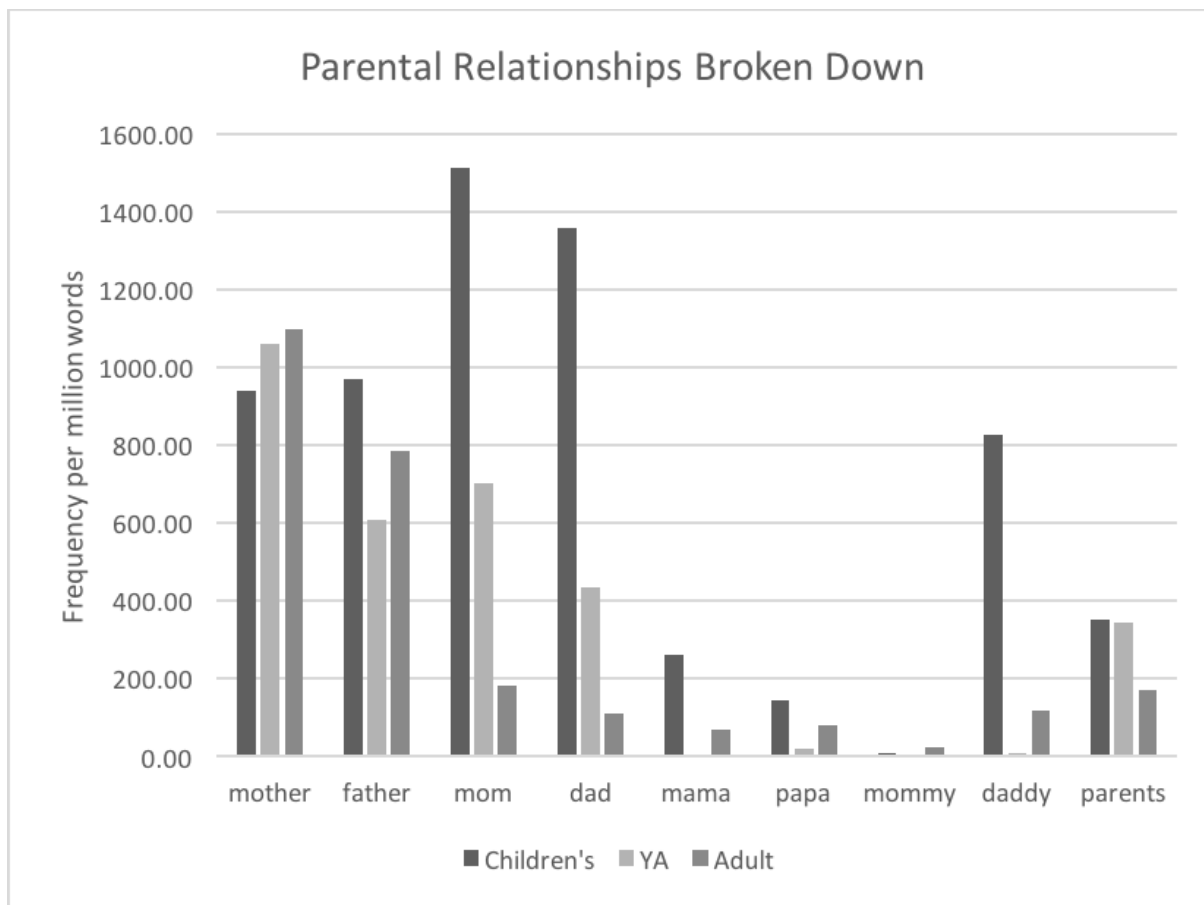
Figure 10: Parental relationships break down

Other familial relationships

We can also see variation between use of other family words. Siblings seem to be discussed most frequently in children's literature. It may be interesting to note that in YA literature, brothers are more common than sisters. The opposite is true in adult fiction. This may be due to the fact that more YA books feature female protagonists and the brother is added to serve as a sort of protective figure or to balance out some of her character traits. Brothers are also more commonly mentioned in children's literature than sisters are.

YA seems to avoid focus on grandparents, while children's literature seems to depict those relationships much more frequently. While differences of formality in address for

parents seemed to vary quite a bit based on age group, formality in address for

grandparents does not seem to follow such clear patterns.

It is somewhat surprising to note that *son* is most common in children's literature.

Logically, it would seem that adult fiction would use the word more, as adults may have

sons, while a child will not. Perhaps this suggests, however, that not only are parents

important in children's literature, but the reciprocal relationship between parent and child

is important.
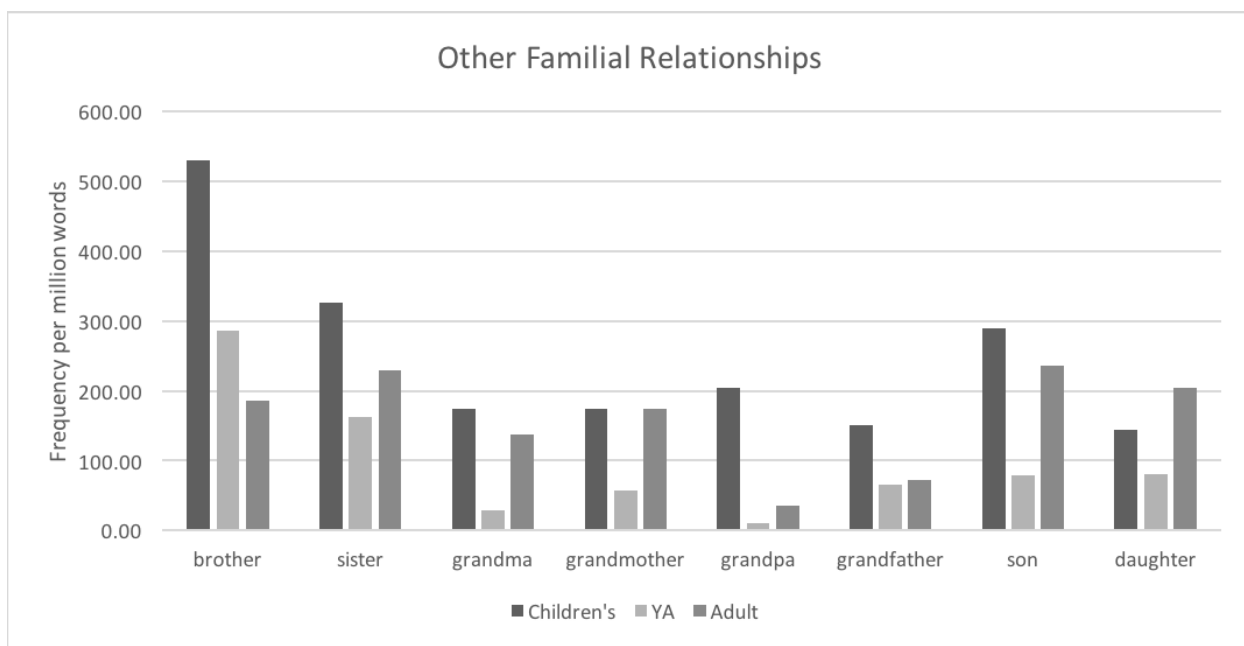
Figure 11: Other familial relationships

CHAPTER FIVE: Conclusions

Summary of findings

The publishing community sees YA literature as distinct from adult and children's literature. The first goal of this study was to see if a distinction could be made linguistically. From the data presented, it seems clear that YA does indeed behave differently than children's and adult literature.

For some features, YA served as a sort of linguistic bridge between children's and adult literature. In other words, the frequency counts for YA would fall between counts for children's and adult. In this sense it connects children's and adults. An example of this would be expletives. They were used not at all in children's, some in YA, and frequently in adult. Conversely, words for parents were most common in children's literature and least common in adults. Here again, YA falls in the middle, bridging the gap between the two.

However, there are also a number of features in which YA seems to act independently of trends in children's and adult. For instance, pronoun use suggests that YA utilizes first person narration more than other registers. High use of body part words, particularly in relation to romance, also seems to correlate strongly with YA literature. On the other hand, YA tends to reject use of animals, perhaps in an attempt to avoid seeming childish.

A second goal of this study was to determine which features demonstrated variation. A number of lexical words showed variation, such as family words and body part words. There were also substantial differences in function words like modals and pronouns. The majority of features examined did demonstrate variation. However, the study was unable to verify the variations between temporal and spatial words found in the

Oxford study. More research should be done in this area to try and clarify the relationship between audience and use of spatial and temporal words.

Finally, I would like to address the issue of whether YA should be considered as a distinct sub-register of fiction. The evidence seems clear that it is distinct from other types of fiction in many features. Furthermore, while it frequently acts as a bridge, it also can act completely differently from either adult or children's books. This suggests that future research should pay more consideration to differences in target audience of fiction texts.

Limitations

As previously mentioned, there are a number of challenges involved in creating a fiction corpus. Texts are highly protected by copyright, making them difficult to obtain. For this study, all texts in the YA corpus had to be scanned and converted to readable text. The process was very time-consuming, which limited the size of the corpus. Corpus size is perhaps the most important limitation in this study. A million-word corpus may not be big enough to study lower frequency items. It also makes the data more subject to skewing by a single text, though the corpus was designed with the hope of avoiding that as much as possible. Expanding the corpus would allow future research to examine more features as well as raise confidence in the results.

One other possible limitation was the use of comparative corpora available. As already mentioned, the non-fiction subset of the corpus was not really utilized because there were no suitable corpora to compare against. Additionally, there were a couple differences between the YA corpus and the two corpora taken from COCA. First, the COCA corpora tagged contractions differently than the YA corpus, which caused some confusion in the keyword lists. Also, the fiction from COCA only consists of first chapters, so it does

not directly match up with the fiction from the YA corpus, which contains text from different parts of the books.

One final limitation was the lack of a dispersion measure. A dispersion measure would require a word to be included in a set number of texts for it to be counted as key. AntConc's current version does not have this feature, however it would have proved useful in creation of keyword lists. Although the corpus was designed to avoid skewing, many high ranking words in the keyword list were proper nouns usually character names. If the dispersion could be set to only include words that appeared in two or more texts, most of these proper nouns would disappear from the keyword lists, making the lists more effective in answering research questions.

Future research

The research done here has begun to establish key differences between YA fiction and fiction for other audiences. However, there is still much that can be done. There are many features not included in this study which could be examined in the future. Perhaps most importantly, future research should attempt to expand the corpora. While there certainly are uses for a corpus of this size, much more could be done with a larger corpus, particularly in regards to examining lower frequency items. For instance, it would be interesting to see if YA books present teens with the type of vocabulary needed to succeed on standardized tests and in college courses.

Variation between sub-registers of YA literature should also be considered in future research. Gardner (2004) has done research examining differences between expository and narrative texts for children. It would also be useful to examine differences between fiction and non-fiction for teens. While a non-fiction subset was created for the corpus, it was

never really utilized as the research ended up focusing more on differences between fiction for different audiences.

Furthermore, it would be interesting to examine differences between genre with in YA fiction. Does YA historical fiction behave differently than YA fantasy? Many readers of YA only read within one or two genres, and would only be exposed to the words in the genres they read. Originally, the corpus for this study was supposed to demonstrate equal representation between different genres. However, many YA books merge or defy standard genre distinctions, making categorization hard. Although such distinctions proved to be beyond the scope of this study, they would be useful for future studies.

It would also be interesting to study how YA has changed over time. While the corpus does contain some older texts (such as *The Outsiders* and *The Chocolate War)*, the majority are more modern. In addition to linguistic change across all registers of English, we would expect to find linguistic evidence that points to trends within YA publishing. Marketplace fads such as vampires and dystopia would likely reveal themselves in a corpus properly suited to tracking change in YA books over time.

Finally, it would be interesting to use corpus-driven methodologies to approach questions about literacy and reading pedagogy. For instance, it would be interesting to analyze the relationship between token count (the number of words in the corpus) and type count (the number of unique spellings in the corpus). Type-token analysis can give insights about lexical density (how many unique words are in a text), which in turn affects readability. It would also be interesting to create a corpus that is organized according to age and then use that as a mechanism for leveling books.

Implications

In recent years, the publishing industry has become increasingly aware of differences between YA texts and texts for younger readers. Based on the linguistic data from this study, it appears that this distinction can be seen at the lexical level as well. As readers grow up, they begin reading more mature texts. We expect texts to evolve both in content and language as they progress to older audiences. However, very little has been done to examine how this progression occurs. Traditionally, when juvenile literature has been included in corpora, it has been grouped together. However, the research presented here suggests that target audience has a real influence on the language used in fiction. Greater attention should be paid to the different branches of juvenile literature, rather than lumping all juvenile fiction together.

As previously mentioned, research has found correlations between reading and vocabulary growth. Exposure to words through reading leads to vocabulary gains, and Nagy, Herman, and Anderson (1985) suggest that this is best accomplished when self motivated readers widely read books of their own selection. With this in mind, it becomes increasingly important to be aware of what type of input children and teens are receiving when they read children's and YA books (which for many young readers are the books they will be most naturally drawn to for self-selected reading). This research has indicated that there are many differences in the language of books readers encounter as they grow older. If vocabulary gains are based on exposure to words, then readers of YA books will be exposed to different words than readers of children's books. However, many of these vocabulary differences are still unexplored, meaning we still do not fully know what type of words younger readers are exposed to, and—by extension—which words they are most

likely to acquire. Corpora can also be useful for gauging whether words encounter with enough helpful context that we can expect readers to derive the meaning of the word if they do not already know it. A greater understanding of the vocabulary found in these texts will allow us to make more informed decisions about how to implement these texts in the classroom.

In conclusion, this study has established some key differences between YA texts and texts for children and adults, identifying YA as a unique linguistic sub-register. However, because so little linguistic research has been done on YA, there are still many aspects which remain unexplored. Future research should seek to expand the corpus size in order to allow examination of more features.

APPENDIX A: Books included in the corpus

Fiction books

Book Title	Author	Genre	Words used in corpus
13 Little Blue Envelopes	Maureen Johnson	Contemporary	12984
A Great and Terrible Beauty	Libba Bray	Fantasy	14989
The Assassin's Blade	Sarah J. Maas	Fantasy	17185
Before I Fall	Lauren Oliver	Fantasy	14699
Blood Promise	Richelle Mead	Fantasy	15651
Boy in the striped Pajamas	John Boyne	Historical	13325
Calling on Dragons	Patrica C. Wrede	Fantasy	14083
The Chocolate War	Robert Cormier	Historical	13546
Coraline	Neil Gaiman	Fantasy	12023
Cress	Marissa Myer	Science Fiction	15946
Dangerous Creatures	Kami Garcia & Margaret Stohl	Fantasy	15128
Daughter of Smoke and Bone	Laini Taylor	Fantasy	15233
Eleanor and Park	Rainbow Rowell	Historical	15265
Evernight	Claudia Gray	Fantasy	15654
Flipped	Wendelin Van Draanen	Contemporary	17118
Gone	Michael Grant	Science Fiction	14261
The Grey Wolf Throne	Cinda Williams Chima	Fantasy	16423
Hex Hall	Rachel Hawkins	Fantasy	12593
The Hunger Games	Suzanne Collins	Science Fiction	17558
I Am Number Four	Pittacus Lore	Science Fiction	13270
If I Stay	Gayle Forman	Contemporary	13229
The Infinite Sea	Rick Yancey	Fantasy	16797
Insurgent	Veronica Roth	Science Fiction	12432
Matched	Ally Condie	Science Fiction	12432
Monster	Walter Dean Meyers	Contemporary	7327
Need	Carrie Jones	Fantasy	15628
The Order of the Phoenix	J.K. Rowling	Fantasy	19306
The Outsiders	S.E. Hinton	Historical	17866

Pretties	Scott Westfield	Science Fiction	13964
The Princess Academy	Shannon Hale	Fantasy	13272
Princess in the Spotlight	Meg Cabot	Contemporary	14200
Rats Saw God	Rob Thomas	Contemporary	16826
Revolution	Jennifer Donnelly	Historical	16197
Shadow and Bone	Leigh Bardugo	Fantasy	13939
Shatter Me	Tahereh Mafi	Science Fiction	14513
Sisterhood of the Travelling Pants	Ann Brashares	Contemporary	13552
Someone Like You	Sarah Dessen	Contemporary	15500
The Icebound Land	John Flanagan	Fantasy	18534
The Luxe	Anna Godberson	Historical	14506
The Thief	Meghan Whalen Turner	Fantasy	20462
Thirteen Reason Why	Jay Asher	Contemporary	12089
Tithe	Holly Black	Fantasy	13401
Twilight	Stephanie Meyer	Fantasy	15650
The Unbecoming of Mara Dyer	Michelle Hodkin	Fantasy	14059
Under the Never Sky	Veronica Rossi	Science Fiction	1468
Unearthly	Cynthia Hand	Fantasy	1450
UnWholly	Neal Shusterman	Science Fiction	18660
Wake	Lisa McMann	Fantasy	11539
Wicked Lovely	Melissa Marr	Fantasy	14488
Willow	Julia Hoban	Contemporary	16763
Wintergirls	Laurie Halse Anderson	Contemporary	13808
Wither	Lauren DeStephano	Fantasy	14658

Appendix B: Keyword Lists

Top 400 Keywords from YA corpus compared against a reference corpus of Children's Fiction

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
1	1879	3499.754	don	40	1024	307.852	hand
2	1500	2922.113	didn	41	182	306.108	kaye
3	848	1674.862	xb	42	260	303.137	janie
4	704	1390.452	xad	43	199	289.828	link
5	685	1327.438	wasn	44	158	285.017	archie
6	604	1179.068	couldn	45	142	280.460	ridley
7	748	1088.986	l	46	140	276.510	aria
8	22824	1025.781	i	47	140	276.510	weren
9	948	1017.622	x	48	137	270.585	aren
10	9498	986.920	she	49	135	266.635	magus
11	478	944.085	doesn	50	286	265.852	harry
12	8764	895.117	her	51	173	255.234	adam
13	352	695.226	xading	52	18954	248.881	and
14	335	661.650	wouldn	53	10590	241.350	it
15	300	592.522	hadn	54	131	239.840	haven
16	279	551.045	isn	55	1149	235.557	still
17	274	541.170	tally	56	119	235.034	obie
18	10626	524.474	he	57	117	231.084	shmuel
19	6872	523.989	my	58	114	225.158	astrid
20	4295	515.839	had	59	114	225.158	scarlett
21	8295	514.599	t	60	284	222.663	r
22	5278	502.734	me	61	112	221.208	shay
23	246	485.868	raisa	62	1342	220.836	even
24	365	479.118	okay	63	111	219.233	tibby
25	240	474.018	coraline	64	111	219.233	xander
26	8568	461.579	that	65	2947	217.867	like
27	8841	460.855	was	66	566	213.059	though
28	2477	447.152	d	67	106	209.358	darry
29	231	444.285	ofthe	68	12263	208.945	of
30	250	436.040	willow	69	1660	208.580	been
31	208	410.815	miri	70	171	208.257	lena
32	1071	401.149	says	71	104	205.408	karou
33	213	399.863	bruno	72	3118	205.252	not
34	3475	384.475	him	73	102	201.457	caine
35	191	377.239	morwen	74	99	195.532	telemain
36	314	369.458	h	75	949	193.835	face
37	238	348.010	eleanor	76	285	191.725	blood
38	1341	333.984	eyes	77	96	189.607	cimorene
39	156	308.111	ginny	78	96	189.607	quinn

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
79	226	180.149	lips	122	63	124.430	haymitch
80	91	179.732	celaena	123	63	124.430	kazul
81	91	179.732	shouldn	124	103	124.390	lucas
82	286	177.847	park	125	424	124.230	moment
83	111	177.493	cinder	126	4096	124.136	as
84	127	172.649	halt	127	1078	121.358	through
85	87	171.831	donia	128	907	120.621	because
86	87	171.831	ky	129	551	120.459	hair
87	87	171.831	xadthing	130	60	118.504	kai
88	102	171.121	bridget	131	60	118.504	nox
89	644	170.926	voice	132	4875	117.103	but
90	132	170.767	diana	133	59	116.529	elody
91	559	167.452	against	134	59	116.529	mendanbar
92	84	165.906	arobynn	135	59	116.529	paisley
93	84	165.906	cress	136	63	115.060	dub
94	119	162.210	aislinn	137	58	114.554	tht
95	82	161.956	britta	138	57	112.579	petrocelli
96	1064	158.119	head	139	359	112.405	seemed
97	1788	158.103	or	140	200	108.864	ii
98	79	156.031	dally	141	63	108.431	dudley
99	79	156.031	starkey	142	101	106.998	tm
100	90	154.047	ht	143	54	106.654	felicity
101	107	152.341	hell	144	70	106.176	adrian
102	115	151.578	lindsay	145	2604	104.796	there
103	75	148.131	peeta	146	126	104.746	perry
104	74	146.155	mal	147	53	104.679	effie
105	73	144.180	ifyou	148	959	104.326	than
106	109	141.141	noah	149	6023	103.637	his
107	130	140.416	carrie	150	52	102.704	vale
108	76	140.363	keenan	151	1082	101.948	say
109	772	139.502	any	152	104	101.853	ing
110	212	139.195	skin	153	51	100.729	thorne
111	70	138.255	soren	154	284	98.770	least
112	87	137.645	henri	155	50	98.754	umbridge
113	69	136.280	beira	156	63	97.854	warner
114	327	135.626	won	157	748	97.004	much
115	73	134.518	hermione	158	49	96.779	evanlyn
116	68	134.305	xa	159	49	96.779	pippa
117	72	132.570	zane	160	1960	95.790	know
118	67	132.330	darkling	161	651	95.018	thing
119	66	130.355	byrne	162	48	94.804	cannot
120	84	127.411	ron	163	47	92.828	brimstone
121	72	125.680	wizards	164	47	92.828	elodie

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
165	80	91.932	carmen	208	39	77.028	hasn
166	134	91.093	connor	209	39	77.028	linden
167	46	90.853	damn	210	39	77.028	liv
168	210	90.761	bit	211	39	77.028	montaigne
169	158	90.525	pain	212	39	77.028	sophos
170	113	90.074	y	213	39	77.028	tobias
171	484	89.965	side	214	39	77.028	zara
172	285	89.425	guy	215	143	76.644	sort
173	53	89.360	corny	216	325	76.411	arms
174	71	89.002	lissa	217	324	76.297	smile
175	45	88.878	shit	218	376	76.245	already
176	152	88.712	sense	219	38	75.053	evernight
177	204	88.684	neck	220	386	73.410	myself
178	301	87.569	breath	221	37	73.078	anza
179	334	87.061	mouth	222	37	73.078	faeries
180	44	86.903	bram	223	41	72.459	cello
181	44	86.903	deparnieux	224	44	72.314	marcus
182	44	86.903	roiben	225	69	71.431	seth
183	44	86.903	zuzana	226	36	71.103	ifhe
184	101	86.467	k	227	36	71.103	olana
185	54	85.944	rolfe	228	36	71.103	strigoi
186	43	84.928	cinna	229	62	70.818	ifi
187	43	84.928	dimitri	230	79	70.659	ally
188	43	84.928	fey	231	323	70.353	arm
189	43	84.928	julianna	232	373	70.033	kind
190	43	84.928	moroi	233	336	69.468	light
191	62	83.679	apparently	234	35	69.128	dauntless
192	42	82.953	ofher	235	35	69.128	gonna
193	42	82.953	ofhis	236	35	69.128	halley
194	42	82.953	scorn	237	35	69.128	nightwalker
195	42	82.953	spink	238	35	69.128	stag
196	2221	82.591	were	239	35	69.128	unwind
197	388	81.981	dark	240	35	69.128	xadtion
198	605	81.364	which	241	2126	68.147	back
199	613	81.084	hands	242	287	68.076	body
200	41	80.978	erak	243	494	67.808	might
201	41	80.978	levana	244	34	67.152	casnoff
202	41	80.978	madrigal	245	34	67.152	demouy
203	41	80.978	peder	246	34	67.152	realms
204	48	79.873	lev	247	34	67.152	tithe
205	91	77.966	gaze	248	58	67.084	trace
206	143	77.928	w	249	75	66.841	slightly
207	39	77.028	bobo	250	155	66.764	god

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
251	56	66.711	jerry	294	537	57.424	behind
252	128	65.547	twenty	295	29	57.277	gretel
253	441	65.409	almost	296	29	57.277	micah
254	1076	65.197	away	297	29	57.277	ponyboy
255	33	65.177	akiva	298	29	57.277	tlu
256	33	65.177	cabel	299	637	57.162	maybe
257	33	65.177	dumbo	300	94	56.662	soda
258	33	65.177	sirius	301	171	55.995	pull
259	1067	64.695	before	302	56	55.942	cassie
260	51	64.530	mara	303	40	55.923	nelson
261	999	64.268	looked	304	59	55.899	lean
262	1215	63.391	around	305	150	55.712	throat
263	32	63.202	amon	306	185	55.439	fact
264	32	63.202	crims	307	28	55.302	esa
265	32	63.202	hl	308	28	55.302	grisha
266	32	63.202	macon	309	28	55.302	penelope
267	381	62.923	both	310	28	55.302	tl
268	135	62.797	f	311	49	55.104	cloak
269	2574	62.372	so	312	1105	54.907	only
270	4757	62.065	at	313	39	54.141	cable
271	41	61.870	zombie	314	39	54.141	gabriel
272	31	61.227	ain	315	27	53.327	assassins
273	31	61.227	edilio	316	27	53.327	felicia
274	31	61.227	grandmere	317	27	53.327	prim
275	31	61.227	renault	318	27	53.327	raquel
276	31	61.227	sampson	319	27	53.327	sounis
277	526	61.196	enough	320	120	53.241	smoke
278	71	60.964	tone	321	1127	53.190	way
279	210	60.770	either	322	90	53.146	expression
280	54	60.499	melinda	323	755	53.097	really
281	955	60.463	something	324	60	53.053	obviously
282	123	59.898	darkness	325	110	52.434	somehow
283	110	59.746	pants	326	40	52.268	mortal
284	171	59.474	stupid	327	259	52.071	probably
285	139	59.457	beneath	328	33	51.709	hc
286	169	59.363	quite	329	33	51.709	ofa
287	30	59.252	archer	330	45	51.354	holland
288	30	59.252	pretties	331	26	51.352	doneval
289	30	59.252	steffan	332	26	51.352	dyer
290	37	59.164	bubbly	333	26	51.352	forcible
291	179	58.248	meant	334	26	51.352	harmon
292	226	57.916	shoulder	335	26	51.352	issie
293	519	57.814	feel	336	26	51.352	linder

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
337	26	51.352	unwinding	369	288	47.490	close
338	26	51.352	vigils	370	24	47.402	amade
339	35	50.933	dis	371	24	47.402	awol
340	35	50.933	naked	372	24	47.402	baghra
341	192	50.793	anyway	373	24	47.402	bastard
342	165	50.571	chest	374	24	47.402	faery
343	686	50.457	knew	375	24	47.402	glamour
344	165	49.768	closer	376	24	47.402	ofthem
345	29	49.440	crap	377	24	47.402	reaping
346	29	49.440	tablets	378	24	47.402	uriah
347	25	49.377	ass	379	85	47.114	g
348	25	49.377	brien	380	423	47.048	trying
349	25	49.377	devyn	381	57	46.693	jaw
350	25	49.377	gemma	382	66	46.657	kissed
351	25	49.377	gotta	383	72	46.583	despite
352	25	49.377	lassiter	384	149	46.041	staring
353	25	49.377	mcgonagall	385	27	45.631	lore
354	25	49.377	nephamael	386	38	45.594	richie
355	25	49.377	ofcourse	387	23	45.427	caster
356	25	49.377	pittacus	388	23	45.427	ifshe
357	25	49.377	socs	389	23	45.427	pol
358	25	49.377	volcra	390	23	45.427	tamron
359	74	49.156	glance	391	524	45.425	turned
360	64	49.104	killer	392	36	45.357	particularly
361	53	49.043	thirteen	393	34	45.307	outsiders
362	116	49.025	kiss	394	182	44.790	beside
363	48	48.295	razor	395	1183	44.480	more
364	51	48.243	chapter	396	44	44.473	district
365	241	48.033	shook	397	207	44.434	actually
366	31	48.002	laurie	398	90	44.172	guards
367	274	47.764	cold	399	51	43.962	fold
368	28	47.534	vernon	400	37	43.916	angela

Top 400 Keywords from YA corpus compared against a reference corpus of Adult fiction

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
1	1879	2474.893	don	42	286	218.159	park
2	22824	2453.235	i	43	142	211.540	ridley
3	1500	1565.451	didn	44	338	204.552	dad
4	848	1263.282	xb	45	135	201.112	magus
5	704	1048.763	xad	46	1612	200.110	re
6	948	1025.963	x	47	140	197.958	aria
7	685	1006.686	wasn	48	140	197.958	weren
8	604	875.518	couldn	49	132	196.643	diana
9	6872	871.126	my	50	173	193.020	adam
10	748	840.192	l	51	199	191.910	link
11	478	688.748	doesn	52	252	188.355	king
12	8295	577.038	t	53	137	185.731	aren
13	352	524.381	xading	54	119	177.277	aislinn
14	335	486.713	wouldn	55	134	174.624	connor
15	300	446.916	hadn	56	117	174.297	shmuel
16	353	409.830	sam	57	112	166.849	shay
17	5278	407.578	me	58	111	165.359	tibby
18	2114	396.953	can	59	111	165.359	xander
19	279	394.441	isn	60	114	159.635	astrid
20	314	391.400	h	61	114	159.635	scarlett
21	260	387.327	janie	62	119	159.475	obie
22	274	387.065	tally	63	106	157.910	darry
23	246	366.471	raisa	64	104	154.931	karou
24	240	357.53	coraline	65	1921	154.330	are
25	238	354.553	eleanor	66	115	153.652	lindsay
26	231	323.687	ofthe	67	102	151.951	bridget
27	208	309.862	miri	68	9461	150.657	you
28	1071	289.303	says	69	101	150.462	tm
29	191	284.536	morwen	70	99	147.482	telemain
30	545	281.884	mom	71	105	146.391	han
31	284	277.931	r	72	96	143.013	cimorene
32	10590	255.012	it	73	91	135.565	celaena
33	182	251.640	kaye	74	91	135.565	shouldn
34	327	245.719	won	75	102	134.761	caine
35	171	243.741	lena	76	90	134.075	ht
36	1974	235.631	m	77	143	133.184	w
37	200	235.317	ii	78	109	132.946	noah
38	213	234.571	bruno	79	111	130.576	cinder
39	156	232.396	ginny	80	87	129.606	donia
40	158	224.532	archie	81	87	129.606	ky
41	250	223.598	willow	82	87	129.606	xadthing

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
83	127	126.617	halt	126	60	89.383	nox
84	84	125.136	arobynn	127	126	89.127	perry
85	755	123.154	really	128	81	88.954	hannah
86	82	122.157	britta	129	84	88.777	ron
87	82	122.157	bryce	130	3346	88.615	we
88	131	121.127	haven	131	59	87.893	elody
89	79	117.688	dally	132	59	87.893	mendanbar
90	79	117.688	starkey	133	58	86.404	tht
91	191	116.660	e	134	1960	85.307	know
92	84	115.551	cress	135	57	84.914	petrocelli
93	76	113.219	keenan	136	63	84.838	dub
94	75	111.729	peeta	137	56	83.424	cassie
95	178	110.998	queen	138	88	82.754	nick
96	115	110.830	steve	139	1082	81.714	say
97	80	109.689	carmen	140	54	80.445	melinda
98	145	108.814	henry	141	54	80.445	rolfe
99	73	108.750	hermione	142	59	79.009	kenny
100	385	107.534	looks	143	59	79.009	lilly
101	71	105.770	lissa	144	53	78.955	effie
102	553	105.486	school	145	194	77.977	c
103	70	104.280	adrian	146	164	77.668	elizabeth
104	70	104.280	soren	147	52	77.465	juli
105	69	102.791	beira	148	52	77.465	vale
106	68	101.301	xa	149	365	76.925	okay
107	145	101.274	j	150	78	76.852	sword
108	2477	100.963	d	151	2947	76.177	like
109	67	99.811	darkling	152	51	75.976	mara
110	85	99.142	g	153	51	75.976	thorne
111	1024	98.840	hand	154	90	75.639	guards
112	66	98.322	byrne	155	925	74.816	ve
113	72	97.981	wizards	156	50	74.486	umbridge
114	72	97.981	zane	157	113	74.342	y
115	79	95.814	ally	158	132	74.013	u
116	103	95.263	lucas	159	49	72.996	evanlyn
117	74	94.318	mal	160	49	72.996	pippa
118	63	93.852	dudley	161	66	72.809	alice
119	63	93.852	haymitch	162	782	72.326	door
120	63	93.852	kazul	163	54	71.736	felicity
121	73	92.882	ifyou	164	171	71.731	stupid
122	62	92.363	ifi	165	48	71.507	bailey
123	3733	91.202	is	166	48	71.507	cannot
124	116	90.256	johnny	167	48	71.507	leon
125	135	89.905	f	168	48	71.507	lev

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
169	1149	70.777	still	212	39	58.099	liv
170	1076	70.487	away	213	39	58.099	montaigne
171	53	70.284	astronaut	214	39	58.099	sophos
172	47	70.017	claire	215	39	58.099	tobias
173	47	70.017	elodie	216	39	58.099	zara
174	76	67.381	princess	217	96	57.827	quinn
175	45	67.037	tucker	218	75	57.377	castle
176	8756	66.151	s	219	44	57.245	daniel
177	50	65.930	peg	220	38	56.609	evernight
178	474	65.910	am	221	43	55.800	edward
179	44	65.548	bram	222	43	55.800	lu
180	44	65.548	deparnieux	223	56	55.271	jerry
181	44	65.548	roiben	224	37	55.120	anza
182	44	65.548	zuzana	225	37	55.120	bethany
183	247	65.532	miss	226	37	55.120	evan
184	94	64.466	soda	227	37	55.120	faeries
185	76	64.274	opens	228	37	55.120	hayden
186	43	64.058	cinna	229	42	54.357	nd
187	43	64.058	dimitri	230	266	54.347	parents
188	43	64.058	fey	231	100	54.171	forest
189	43	64.058	jenna	232	4875	54.008	but
190	43	64.058	julianna	233	49	53.921	cloak
191	43	64.058	moroi	234	117	53.712	magic
192	65	63.430	shakes	235	36	53.630	ifhe
193	101	63.260	k	236	36	53.630	janet
194	42	62.568	ofher	237	36	53.630	olana
195	42	62.568	ofhis	238	36	53.630	strigoi
196	42	62.568	spink	239	88	53.355	totally
197	47	61.583	brimstone	240	86	53.292	weird
198	3237	61.344	what	241	106	53.028	smiles
199	41	61.079	erak	242	41	52.915	cello
200	41	61.079	levana	243	41	52.915	zombie
201	41	61.079	madrigal	244	71	52.636	spell
202	41	61.079	peder	245	59	52.267	reaches
203	157	60.602	turns	246	1059	52.235	will
204	651	59.479	thing	247	35	52.140	briggs
205	53	59.422	corny	248	35	52.140	dauntless
206	59	59.240	paisley	249	35	52.140	gonna
207	45	58.690	ti	250	35	52.140	halley
208	39	58.099	bobo	251	35	52.140	nightwalker
209	39	58.099	hasn	252	35	52.140	unwind
210	39	58.099	kaz	253	35	52.140	xadtion
211	39	58.099	linden	254	2604	52.070	there

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
255	34	50.650	casnoff	298	41	43.038	shelly
256	34	50.650	demouy	299	34	42.857	realms
257	1239	50.628	think	300	34	42.857	tithe
258	33	49.161	akiva	301	1191	42.359	right
259	33	49.161	cabel	302	37	41.918	bubbly
260	33	49.161	hc	303	96	41.778	th
261	33	49.161	ofa	304	28	41.712	esa
262	33	49.161	sirius	305	28	41.712	grisha
263	386	48.822	myself	306	28	41.712	tl
264	457	48.342	someone	307	613	41.699	hands
265	323	48.217	arm	308	85	40.914	starts
266	32	47.671	amon	309	36	40.535	ne
267	32	47.671	crims	310	27	40.222	assassins
268	32	47.671	hl	311	27	40.222	emma
269	32	47.671	lars	312	27	40.222	felicia
270	32	47.671	missy	313	27	40.222	pixies
271	32	47.671	pixie	314	27	40.222	raquel
272	55	47.235	caesar	315	27	40.222	sounis
273	44	47.098	marcus	316	439	40.174	hear
274	57	46.726	nods	317	32	39.997	lunar
275	1019	46.657	look	318	44	39.692	holly
276	1347	46.385	see	319	41	39.234	guitar
277	285	46.363	sorry	320	305	39.199	everyone
278	31	46.181	brooke	321	1078	39.027	through
279	31	46.181	darla	322	59	38.836	academy
280	31	46.181	edilio	323	26	38.733	doneval
281	31	46.181	grandmere	324	26	38.733	dyer
282	31	46.181	renault	325	26	38.733	forcible
283	63	45.807	warner	326	26	38.733	harmon
284	394	45.675	stop	327	26	38.733	issie
285	58	45.128	gym	328	26	38.733	linder
286	284	45.066	happened	329	26	38.733	unwinding
287	30	44.692	archer	330	26	38.733	vigils
288	30	44.692	steffan	331	86	38.624	pulls
289	42	44.388	scorn	332	164	38.613	class
290	445	43.870	mean	333	31	38.569	ain
291	52	43.504	teddy	334	31	38.569	lynn
292	29	43.202	gretel	335	31	38.569	sampson
293	29	43.202	lt	336	45	37.820	holland
294	29	43.202	micah	337	37	37.673	angela
295	29	43.202	ponyboy	338	217	37.428	friends
296	29	43.202	tlu	339	47	37.379	closes
297	1077	43.178	going	340	25	37.243	devyn

Rank	Frequency	Keyness	Keyword	Rank	Frequency	Keyness	Keyword
341	25	37.243	gemma	371	32	35.034	macon
342	25	37.243	gotta	372	3704	35.031	said
343	25	37.243	jeffrey	373	35	35.014	stag
344	25	37.243	lassiter	374	1215	34.938	around
345	25	37.243	mcgonagall	375	96	34.688	professor
346	25	37.243	nephamael	376	40	34.634	mortal
347	25	37.243	nesbitt	377	28	34.300	knight
348	25	37.243	ofcourse	378	28	34.300	vernon
349	25	37.243	pittacus	379	23	34.264	bianca
350	25	37.243	socs	380	23	34.264	ifshe
351	25	37.243	volcra	381	23	34.264	juvies
352	30	37.144	grams	382	23	34.264	kenji
353	30	37.144	pretties	383	23	34.264	tamron
354	285	36.925	blood	384	486	34.121	need
355	44	36.583	dancer	385	2154	34.116	no
356	150	36.433	human	386	71	33.979	prison
357	66	36.053	puts	387	141	33.939	anymore
358	193	36.014	knows	388	76	33.914	grade
359	48	35.877	leans	389	34	33.691	vaughn
360	24	35.753	amade	390	31	33.668	laurie
361	24	35.753	baghra	391	39	33.382	gabriel
362	24	35.753	faery	392	753	33.244	tell
363	24	35.753	guardians	393	27	32.882	carter
364	24	35.753	ng	394	22	32.774	archaniz
365	24	35.753	ofthem	395	22	32.774	cathy
366	24	35.753	ot	396	22	32.774	highness
367	24	35.753	reaping	397	22	32.774	juvey
368	24	35.753	uriah	398	22	32.774	ls
369	151	35.231	asks	399	22	32.774	malfoy
370	436	35.063	oh	400	22	32.774	md

APPENDIX C: Tables

Table 1: Modals

	Childrens	YA	Adult
can	2273.1	2732.1	1038.4
could	2149.6	2406.4	2176.3
shall	92.8	34.9	30.8
should	696.6	620.3	465.3
will	2339.1	1368.6	760.4
would	2082.8	2517.5	2553.8
may	383.6	161.5	822.0
might	376.7	638.4	427.3
must	721.9	392.9	401.0
total	11116.2	10872.7	8675.4

Table 2: Body part words

	Childrens	YA	Adult
eyes	820.9	1733.1	1091.8
hand	572.3	1323.4	632.8
face	631.4	1226.5	845.5
blood	87.5	368.3	161.1
lips	56.8	292.1	161.1
head	791.7	1375.1	901.7
skin	66.7	274.0	169.3
hair	356.7	712.1	431.8
arm	209.4	417.4	173.8

Table 3: Pronouns

	Children's	YA	Adult
I	22167.16	29497.10	18086.83
you	14416.52	12227.13	9851.39
he	10169.50	13732.75	13518.73
she	7818.15	12274.95	12654.17
we	4964.31	4324.28	3380.36

Table 4: Time words

	Children's	YA	Adult
late	210.20	211.95	225.42
later	663.60	273.98	366.64
before	985.80	1378.96	1286.42
after	1246.64	1050.70	1287.33
during	238.59	115.02	243.52
soon	441.89	290.78	295.13
now	1650.93	1904.96	1562.53
often	118.14	78.83	194.64
usually	75.18	107.27	121.31
never	1036.44	1168.30	1226.67

Table 5: Spatial words

	Children's	YA	Adult
bottom	112.77	105.97	106.82
edge	83.62	171.89	171.10
gap	11.51	19.39	20.82
shape	54.47	59.45	64.28
hole	135.79	72.37	105.01
side	333.72	625.51	586.63
slope	10.74	14.22	37.12
top	255.47	231.33	276.11

Table 6: Expletives

	Children's	YA	Adult
damn	0.00	58.16	162.05
hell	7.67	134.41	245.33
shit	0.00	58.16	148.47
bitch	0.00	28.43	72.42
bastard	0.00	31.02	35.31
ass	0.00	32.31	93.25
fuck	0.00	18.09	51.60

Table 7: Animal Words

	Children's	YA	Adult
dog	301.49	59.45	168.38
cat	239.35	161.55	82.38
bird	140.39	25.85	456.27
horse	214.04	107.27	119.50
fish	178.75	46.53	85.10
bear	142.69	58.16	96.87
snake	88.22	25.85	19.92
pig	54.47	12.92	62.47
cow	83.62	3.88	53.41
lion	82.85	11.63	31.69
wolf	51.40	105.97	388.37

Table 8: Spatial words

	Children's	YA	Adult
bottom	112.77	105.97	106.82
edge	83.62	171.89	171.10
gap	11.51	19.39	20.82
shape	54.47	59.45	64.28
hole	135.79	72.37	105.01
side	333.72	625.51	586.63
slope	10.74	14.22	37.12
top	255.47	231.33	276.11

Table 9: Parental relationships

	Children's	YA	Adult
Both parents	5640.18	3194.74	2573.75
Mother	2726.50	1777.01	1376.95
Father	2561.56	1073.96	1023.89

Table 10: Parental relationships break down

	Children's	YA	Adult
mother	939.77	1062.33	1099.93
father	972.76	607.41	784.89
mom	1515.15	704.34	182.87
dad	1360.18	436.82	110.45
mama	263.14	5.17	68.80
papa	145.76	20.68	81.48
mommy	8.44	5.17	25.35
daddy	828.54	9.05	120.40
parents	352.12	343.77	172.91

Table 11: Other familial relationships

	Children's	YA	Adult
brother	530.88	286.91	186.49
sister	326.81	162.84	229.94
grandma	174.91	29.72	137.60
grandmother	174.91	58.16	173.82
grandpa	204.07	10.34	36.21
grandfather	151.13	65.91	72.42
son	289.22	78.83	235.38
daughter	144.23	80.13	204.60

REFERENCES

Beck, I. L. (2002). In McKeown M. G., Kucan L. (Eds.), *Bringing words to life* Guilford Press, New York.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory, 8*(1). Retrieved February 1, 2016.

Biber, D., & Finegan, E. (1994). *Sociolinguistic perspectives on register*. New York: Oxford University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Brown, A., & Crowe, C. (2013). Ball don't lie: connecting adolescents, sports, and literature. *The ALAN Review, 41*(1), 76. Retrieved January 30, 2016.

Brown, D. W. (2011, August 1). How Young Adult Fiction Came of Age. *The Atantic*. Retrieved February 8, 2016.

Davies, M. (2008-) *The Corpus of Contemporary American English: 520 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Dossena, M. (2012). Vocative and diminutive forms in Robert Louis Stevenson's fiction: A corpus-based study. *International Journal of English Studies, 12*(2). Retrieved February 1, 2016.

Dubin, F., & Olshtain, E. (1993). Predicting word meanings from contextual clues: Evidence from L1 readers. In T. Huckin, M. Haynes, & J. Coady (Eds.), Second language reading and vocabulary learning (pp. 181–202). Norwood, NJ: Ablex Publishing Corporation.

Durand, E. S. (2013). Forging Global Perspectives Through Post–colonial Young Adult Literature. *The ALAN Review, 40*(2). Retrieved January 30, 2016.

Evans, M. A., & Saint-aubin, J. (2013). Vocabulary acquisition without adult explanations in repeated shared book reading: An eye movement study. *Journal of Educational Psychology, 105*, 596-608.

Fischer-Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's Pride and Prejudice: A corpus-stylistic analysis. *IJCL International Journal of Corpus Linguistics, 14*(4), 492-523.

Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied Linguistics, 25*, 1-37.

Gardner, D. (2008). Vocabulary recycling in children's authentic reading materials: A corpus-based investigation of narrow reading. *Reading in a Foreign Language, 20*, 92-122.

Gilmore, N. (2015, August 19). New York Times' Changes Children's Bestsellers Lists. Retrieved February 8, 2016.

Gonzalez, J. E., Pollard-Durodola, S., Simmons, D. C., Taylor, A. B., Davis, M. J., Fogarty, M., & Simmons, L. (2014). Enhancing preschool children's vocabulary: Effects of teacher talk before, during and after shared reading. *Early Childhood Research Quarterly, 29*, 214-226.

Gurdon, M. (2013, November 1). Young Adult Literature's Sorry State. *USA TODAY*, 40-43. Retrieved January 30, 2016.

Grabowski, Ł. (2013). Register Variation Across English Pharmaceutical Texts: A Corpus-driven Study of Keywords, Lexical Bundles and Phrase Frames in Patient Information Leaflets and Summaries of Product Characteristics. *Procedia - Social and Behavioral Sciences, 95*, 391-401. Retrieved February 17, 2016.

Hayn, J. A., & Kaplan, J. S. (2012). *Teaching young adult literature today: Insights, considerations, and perspectives for the classroom teacher*. Lanham: Rowman & Littlefield.

Herz, S., & Gallo, D. (1996). *From Hinton to Hamlet: building bridges between young adult literature and the classics*. Westport, Connecticut: Greenwood Press.

Kennedy, G. D. (1998). *An introduction to corpus linguistics*. London: Longman.

Lamb, W. (2008). *Scottish Gaelic speech and writing: Register variation in an endangered language*. Belfast: Cló Ollscoil na Banríona.

Mahlberg, M. (2010). Corpus Linguistics and the Study of Nineteenth-Century Fiction. *Journal of Victorian Culture, 15*(2), 292-298.

Mahlberg, M. (2012). Corpus stylistics--Dickens, text-drivenness and the fictional world. *Essays and Studies, 65*. Retrieved February 1, 2016.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*, 233-253.

Pacheco, M. B., & Goodwin, A. P. (2013). Putting two and two together: Middle school students' morphological problem-solving strategies for unknown words. *Journal of Adolescent & Adult Literacy, 56*, 541-553.

Parodi, G. (2013). Genre organization in specialized discourse: Disciplinary variation across university textbooks. *Discourse Studies,16*(1), 65-87.

Sardinha, T. B., Kauffmann, C., & Acunzo, C. M. (2014). A multi-dimensional analysis of register variation in Brazilian Portuguese. *Corpora, 9*(2), 239-271.

Siepmann, D. (2015). A corpus-based investigation into key words and key patterns in post-war fiction. *FOL Functions of Language, 22*(3), 362-399.

Stuart, M., Dixon, M., Masterson, J., & Gray, B. (2003). Children's early reading vocabulary: Description and word frequency lists. *British Journal of Educational Psychology, 73*, 585-598.

Wild, K., Kilgarriff, A., & Tugwell, D. (2013). The oxford Children's corpus: Using a Children's corpus in lexicography. *International Journal of Lexicography, 26*, 190-218.

Zhang, Z. (2012). A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory, 8*(1).