How accessible are idioms to children? A Corpus Study Investigating Idiom Frequency in English Juvenile Fiction

Bachelor's Project Thesis

Amanda Komulainen s4242807, a.p.komulainen@student.rug.nl,
Supervisors: Stephen Jones & Jacolien van Rij

1 Introduction

Idiomatic expressions present themselves frequently in everyday language (Nattinger & DeCarrico, 1992). An idiom is commonly defined as an expression whose meaning cannot be derived from the meaning of its comprising words. Examples of such expressions are *to keep an eye on*, meaning to watch over something; or *at the end of the day*, meaning when everything is taken into consideration. As a subset of multiword expressions, idioms tend to have a fixed configuration of words. Multiword expressions function to convey common and recurrent concepts during discourse (Martinez & Schmitt, 2012), which may be the reason why the expression has to be recognizable through its consistent word arrangement.

Certain idioms can have two semantic representations; literal and non-literal (figurative), which may cause ambiguity in deriving the intended meaning of the idiom. For example, *at the end of the day* can literally refer to the time towards the end of a day, but can also figuratively mean when all things are taken into consideration. Hence, in addition to learning a fixed configuration of words, language learners must also learn to derive the intended meaning of an idiom. Understanding idioms can be a challenging task for language learners and the ability to process idioms is therefore an indication of fluency in language (Pawley & Syder, 1983).

As a fundamental part of language proficiency, children must also learn to process idioms in their journey of language development. Due to the additional complexity involved in processing idioms, research have been showcasing interest towards how children develop idiom understanding. For example, Lodge & Leach (1975); Levorato & Cacciari (1992) observed that children around the age of 6 tend to interpret idiomatic phrases in a literal way, suggesting that literal meanings are acquired by children prior to learning non-literal meanings (Chafe, 1970).

The ability to interpret idioms in a non-literal sense has been found to emerge in children around the ages of 8 and 9 (Spector, 1996; Nippold & Martin, 1989), which then expands until the end of childhood (around 18 years old) and continues throughout adulthood (Sprenger et al., 2019; Nippold, 2006). These studies suggest that idiom understanding is a continuously developing knowledge, that even proficient language users like adults learn throughout their life.

There are different viewpoints to how this idiom knowledge develops, particularly in children. One of them follows the idea that familiarity, or frequency of exposure, is associated with proper idiom understanding. In other words, children would understand idioms that they have encountered frequently in their linguistic environment (Nippold & Taylor, 2002; Sprenger et al., 2019; Ezell & Goldstein, 1991).

On the other hand, studies by Levorato & Cacciari (1992, 1995) suggest that children need to develop figurative competence in order to understand idioms. Figurative competence allows a child to detect an expression as a potentially non-literal one, which consequently activates the need to utilize contextual cues and prior knowledge to determine the intended meaning of the expression. Figurative competence has been observed to surface in children around the age of 8, which is inline with the observations of Nippold & Martin (1989); Spector (1996), who found children of this age to display idiom understanding.

Levorato & Cacciari (1992, 1995) argue that familiarity (i.e. frequency of exposure) alone does not necessarily improve idiom understanding, unless the child possesses figurative competence. Familiarity may allow the child to recognize a string as an idiom, but it does not always help in interpreting the figurative meaning of the idiom at hand.

In either of the viewpoints, exposure to idioms is a part of the investigation on children's idiom understanding. One viewpoint suggests that frequent exposure can explain how children acquire and understand idioms, while the other suggests that sufficient language development must be present alongside of exposure. Regardless of which viewpoint one subscribes to, it can be concluded that for children around the age of 8 and above (at least when figurative competence is achieved), exposure plays a role in developing idiom knowledge.

Frequency of exposure to idioms is often measured through familiarity ratings (Levorato & Cacciari, 1992; Sprenger et al., 2019; Gibbs, 1987; Lodge & Leach, 1975; Nippold & Taylor, 2002). The children (or teachers, such in Levorato & Cacciari, 1992) would rate their familiarity on certain idioms listed by the researchers. Familiarity ratings allow an efficient measure on how frequently present an idiom is in one's environment (Libben & Titone, 2008; Tabossi et al., 2011; Titone & Connine, 1994). It takes into account subjective ratings of familiarity, on top of incorporating how well an idiom is understood by the individual.

However, using familiarity as an indicator of idiom exposure operates under the premise that idioms would always be stored in memory. This premise does not take into account that other factors may influence which idioms are remembered and not, such as individual differences in idiom knowledge (Nordmann et al., 2014). For example, individual differences may occur in rating the decomposability of an idiom, which is the degree of how closely related the non-literal meaning of an idiom is to the literal meaning of its composing words. As such, familiarity to idioms alone may not capture the extent of idiom exposure.

An additional way to measure frequency of exposure is through a bottom-up approach that looks at idiom presence directly through the children's linguistic environment. This approach can be done by performing a corpus analysis, which entails inspecting sampled data of naturally-occurring language. The analysis is done by retrieving frequent phrases from the data and scanning these phrases for idioms. A bottom-up approach, as opposed to a top-down one, does not search for specific idioms and allows frequent idiom to surface from the environment observed.

Accessing the use of idioms in naturally-occurring language can give insight to an objective, count-based frequency of exposure instead of based on subjective familiarity. This is especially important in cases where familiarity ratings are provided by teachers (such in Levorato & Cacciari, 1992), because adult familiarity ratings are generally different compared to children (Sprenger et al., 2019).

Analyzing corpora allows researchers to uncover patterns in language that would not be revealed by intuition alone (Biber et al., 1994). Therefore, taking into account what idioms often occur naturally around children, in addition to the children's familiarity ratings, would strengthen the measure of children's idiom exposure.

Corpus analyses on idiom frequency in children's environments are generally scarce. According to van Rij et al. (2023), a potential reason why corpus studies are less common is due to the difficulty in finding idioms through the corpus. The difficulty lies in attempting to capture the different variations of an idiom, such as variations in syntax or word use, and insertions or modification of adverbs or adjectives. Additionally, there is a need for a manual inspection to confirm whether an idiom was intended in a non-literal sense or not.

The generalizability of corpus studies is also limited to the data in the corpus. Even though observations made from corpus data reflect real language use to an extent, it does not encapsulate the entirety of language because the data remains a sample of real-life language. In addition, the frequency threshold chosen for searching frequent idioms affects which idioms are retrieved. In other words, an infrequent idiom that falls below the chosen frequency threshold might not be captured by this approach, while this idiom is still nonetheless present in the language environment. Regardless of these limitations, corpus studies still give important insight into how language is used in real life.

Recent bottom-up studies that looked at idioms through corpus analyses have commonly been using adult corpora. For example, Martinez & Schmitt (2012) assembled a list of most frequent multiword expressions (including idioms) from the *British National Corpus* and Liu (2003) studied idiom frequency in *English as a Second Language* text-books. Observations made from general adult corpora, which includes language used by all ages, does reflect parts of idiom exposure in children. However, it does not capture children's idiom exposure through language targeted primarily towards children.

There are existing studies that have investigated idiom presence in language around children, more specifically in the context of education. For example, as cited in Abkarian et al. (1992), Nippold (1991) found that in three reading programs of 3rd to 8th graders, an idiom occurred in 6.7% of all sentences; additionally, Lazar et al. (1989) found that an idiom occurred at least once in 11% of teacher's classroom utterances. A more recent corpus study by van Rij et al. (2023) also investigated Dutch idiom exposure on children using a corpus of children's literature in Dutch. Overall, there is not much known about English idiom presence in children's language environments outside education.

The current study will therefore perform a bottom-up corpus analysis on children literature using an American English juvenile fiction corpus. Studying children's literature allows insight into content that children may consume outside the previously studied educational environment. Aforementioned studies have shown that children around the age of 8 and above (until 18, which marks the end of childhood) display sufficient language development and possess the ability to figuratively interpret idioms. This observation holds whether one believes in the exposure-viewpoint or the figurative competence-viewpoint described before. Juvenile fiction conveniently targets children between 7 and 18-years old, which is why it was chosen for the present study.

The corpus is provided by COCA: The Corpus of Contemporary American English (Davies, 2008-). COCA is selected as it is one of the most widely used corpus, in addition to being the largest, balanced corpus of American English.

The juvenile fiction corpus contains around 2000 entries of fictional works targeted towards young audiences, with around 3 million words in total. The fictional works consist of stories from books or children magazines. There was no information on the target age range of the juvenile fiction works included the corpus; hence, the juvenile fiction is assumed to target children under the juvenile age range of 7 to 18-years old.

To assess children's idiom exposure through literature, this investigation is made in pursuit of the research question: how frequent are idiomatic expressions present in English juvenile fiction? In the context of this study, frequency refers to occurrence counts.

As the mentioned studies have shown that children under the juvenile age category (7 to 18-years old) do possess the ability to understand idioms, it may imply that children have exposure to certain idioms in their linguistic environment. Juvenile fiction, as a part of that environment, is therefore hypothesized to contain idioms.

In order to answer this research question, this study aims to capture frequent co-occurring words from the juvenile fiction corpus. Frequent co-occurring words may indicate that the sequence of words is a fixed, multiword expression, such as an idiom.

2 Method

The methodology used in the current project includes two steps: extracting frequently co-occurring words with N-grams (Subsection 2.1), followed by a manual inspection of the N-grams to determine the idiomaticity of the N-grams (Subsection 2.2). Before going into the specific methodology, the reasoning behind using N-grams to retrieve possible idiomatic phrases is explained, as well as how variations in idioms are taken into account.

N-grams consist of N neighboring words, which means that a frequent N-gram represents a string of words that occur together frequently. This frequent string of words can potentially be an idiomatic string, which is the intuition behind using N-grams to capture idioms from the data. The same method was applied in a study by Martinez & Schmitt (2012) to extract multiword expressions, thus inspiring the method in the current study.

As mentioned in the Introduction, searching for idioms in a corpus can be challenging due to possible variations in an idiom. Certain variations, such as insertions or modifications of adverbs and adjectives, can still be captured by exploring a wide range of N degrees so that the inserted word will be captured by longer N-grams. Lexical or syntactic variations would have to be discovered during the manual inspection step.

Syntactic variation involving tense variation can be minimized by normalizing the tense of the words to the same, standard tense. For this reason, the lemmatized version of the corpus data was used. Lemmatizing consists of transforming words to its standard, canonical form (e.g., *kicked, kicking* to *kick*). If all words are lemmatized, variations in the tense are diminished. Using the original text was still necessary since the lemmatized text is often incoherent, for example "I am going to" becomes "I be go to".

2.1 Extraction Program

In this Subsection, the steps taken in creating the N-gram extraction code is described from the initial steps to the last. These steps include data transformation and lemmatizing, usage of built-in functions from the *Python* library, extracting and storing the N-grams, and finally retrieving the most frequent N-grams out of the stored N-grams. Programming decisions, such as setting parameter values, are listed and explained. The full code can be accessed via github.com/amandadotkom/N-gram-extractor for more detailed information. The corpus data is only available with permitted access and is therefore not included in the repository.

To handle the textual data from the corpus, a code specific to this study was implemented in *Python 3.11* (Van Rossum & Drake, 2009). Using code specific to the present study eased the data manipulation process and ensured that the format of the results is as desired. Additionally, it allowed the present study to use the lemmas from the COCA corpus itself; ensuring that the lemma used for each word in the corpus was consistent.

Before extracting the N-grams, the text from the corpus was cleaned by removing characters, such as punctuation and numbers, that would not be informative in terms of finding idioms. All the periods were kept in the data to mark the end of a sentence. Question marks and exclamation marks were substituted by periods and therefore also marked the end of sentences. All letters were also transformed to lowercase to improve the visualization of the text.

The data was separated and converted into lemmas by using the lemma chart from COCA, which resulted in two datasets: the original data and the lemmatized version of the data. The lemmatized version of the data was realized by taking the lemmas from the chart and converting them from a table format into a text format. The order of the lemmas in the table corresponded to the position of its original word in the original data, so the lemmatized data matches the order of words in the original data. From here onwards, every step that was done for the original data was also done for the lemmatized data.

After the data was cleaned and transformed, the N-gram extraction steps were taken. The initial step was to store all N-grams with varying *N* degrees from within each sentence in the data. To do this, a function *ngrams()* imported from the *nltk* library (Bird et al., 2009) was used. The *ngrams()* function takes in a list of sentences and the desired N-gram length or degree *N*. For each sentence in the list, the function extracts N-grams of the given *N* degree and stores them. As an example, 4-grams from the sentence *mary jane woke up excited* would be *mary jane woke up* and *jane woke up excited.* It should be noted that the N-grams do not cross sentences, so any idiomatic expressions that were captured spanned within one sentence.

The range of N-grams searched in the data was between 3- to 10-grams to account for varying idiom lengths (for example, due to insertions of adjectives or adverbs). There were not a lot of 10-grams that occurred above the frequency threshold chosen, which is why N values higher than 10 were not explored. The relationship of the frequency threshold and the N-grams will be explained shortly. Any potential idioms that were used in sentences longer than 10 tokens were therefore missed.

From the previous step, 3- to 10-grams from each sentence in the data were stored in lists. Each degree of *N* has its separate list. The next step was to filter the stored N-grams from N-grams containing irrelevant phrases and from infrequent N-grams that fall below a selected frequency threshold.

Certain N-grams containing irrelevant characters tended to saturate the list of N-grams. To shorten the list for manual inspection, the irrelevant N-grams were pruned. These N-grams contained pronouns such as *i, me, you, we, us* and contractions like *would n't, did n't* and *do n't.*

Any potential idioms that contained pronouns would be captured by the shorter N-grams. For example, an idiom in a sentence *i am going to hit the sack*, will not be captured by a 7-gram or higher,

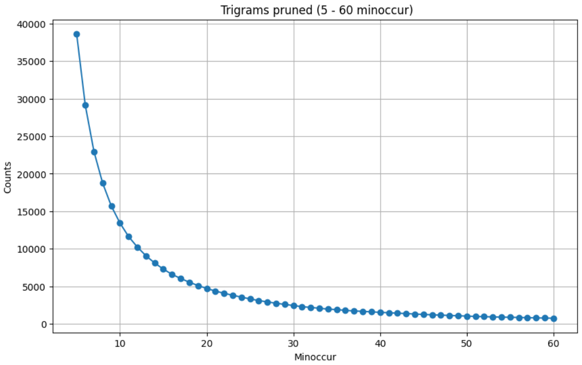

Figure 2.1: Plot of *min_occur* values between 5 to 60 and the resulting number of pruned 3-grams retrieved

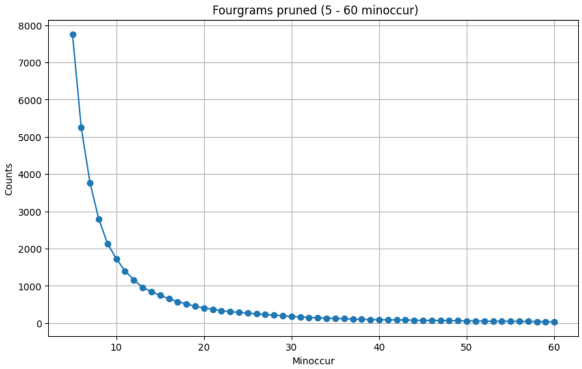

Figure 2.2: Plot of *min_occur* values between 5 to 60 and the resulting number of pruned 4-grams retrieved

but will be captured by 6-grams as *am going to hit the sack*, or by 5-grams as *going to hit the sack*. However, a drawback occurs if a certain idiom with pronouns occurred less times than the *min_occur* value of the shorter N-grams. The lower degree N-grams (3- to 5-grams) are generally too short to include both pronouns and an idiomatic phrase at the same time. So in most cases, pruning the pronouns in N-grams would not cause idioms to be overlooked

The contractions are expanded in the lemmatized data (e.g., *did n't* becomes *did not*), so idioms that did have contractions would be discovered through the lemmatized N-grams. Other non-informative, frequent words in the data such as *of the, narrator, parantheses*, and so on were also pruned. These frequent words were discovered after the initial pruning stage (with pronouns and contractions) and were pruned after the fact. The rest of the pruned characters can be seen in the code. This step was done to ease the inspection process by further reducing the number of N-grams in the lists produced.

For the purpose of filtering infrequent N-grams, the *Counter* tool was used from the *Collections* module within *Python*. The *Counter* provides a function *most_common* that can return the most frequent items (in this case N-grams) from a list and their corresponding occurrence counts.

The frequency threshold that defines which N-grams are sufficiently frequent from the N-gram lists is set by a *min_occur* parameter, which was manually defined and is not a part of the *Counter* tool. If the N-gram occurred at least *min_occur*-times, it was included in the final frequent N-gram list. A high value of *min_occur* results in fewer, high frequency N-grams and vice versa. The *min_occur* parameter was set to different values, depending on the degree of the N-gram. The frequency threshold selection is important as it can cause the idioms below the frequency threshold to be overlooked. An ideal *min_occur* value is one that results in as many frequently occurring N-grams as possible (so as to not overlook potential idioms), while keeping the list of frequent N-grams short enough for a feasible manual inspection.

For N-grams between 6- and 10-grams, the lowest possible *min_occur* value was chosen, which was set to be 3 minimum occurrences. Because the high degree N-grams resulted in relatively fewer number of N-grams, having a low *min_occur* value did not result in a large number of N-grams and was feasible for the manual inspection step.

To select an ideal value for the lower degree N-grams (3- to 5-grams), different *min_occur* values were explored by plotting different values of *min_occur* against the corresponding number of resulting N-grams.

For 3-grams, 15 minimum occurrences was selected as the optimal *min_occur* value, as it lies approximately on the elbow of the plot as seen in Figure 2.1. For the same reason, 14 minimum occurrences were chosen for 4-grams as can be observed in Figure 2.2. The explored *min_occur* values were between 5 to 60. Lower *min_occur* values below 5 would result in large numbers of N-grams beyond

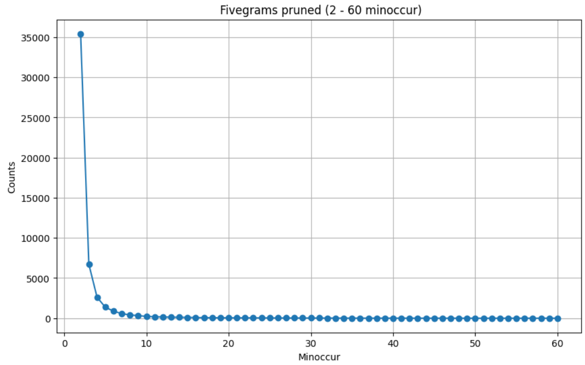

Figure 2.3: Plot of *min_occur* values between 2 to 60 and the corresponding number of 5-grams retrieved.

the manual inspection capabilities of this study.

For 5-grams, a minimum occurrence of 5 was chosen from exploring *min_occur* values between 2 and 60 (Figure 2.3). The overall number of 5-grams retrieved from the corpus was not as large as 3- or 4-grams, which is why the lowest *min_occur* value explored was 2 minimum occurrences.

After filtering the stored N-grams using the corresponding *min_occur* values, 10 lists from the original data and an additional 10 lists from the lemmatized data were made. These lists contain the most frequent N-grams for each *N*. The selected *min_occur* values for each *N* degree can be observed in Table 2.1.

N	*Min_occur*
3	15
4	14
5	5
6	3
7	3
8	3
9	3
10	3

Table 2.1: *Min_occur* value for each degree of *N*

2.2 Manual Inspection

To find idioms in the generated frequent N-grams lists, each N-gram was categorized as idiomatic and non-idiomatic. The idiomatic N-grams found in the original and lemmatized list were combined.

Judgement of the N-gram's idiomaticity was partly based on the judgement of a non-native English speaker, a native British English speaker, the use of an American English idiom dictionary; *The American Heritage dictionary of idioms* (Ammer, 2013), and an online idiom dictionary (Farlex, 2024). During this step, the judgement on idiomaticity was done in isolation; meaning the N-grams were judged without its context.

Relying on a native speaker's opinion resulted in reliable idiomaticity judgements, even though the idioms recognized are limited to the speaker's knowledge. For example, certain idioms are so often used in a language that it becomes a part of the standard lexicon; thus a native speaker might not recognize the idiom as one.

Each potentially idiomatic N-gram judged was also cross-checked using the idiom dictionaries; however, existing idioms can shift in meaning and new idioms may surface with time, which cannot be swiftly updated in dictionaries. To confirm the idiomaticity of potentially idiomatic N-grams, the N-grams that made it to the final idiom list had to be deemed idiomatic by both the native speaker and the idiom dictionaries.

After filtering the frequent N-grams list down to idiomatic N-grams (from now on referred to as an idiom), each idiom was inspected in its original context to determine if it was used figuratively or literally. The contexts where the idioms appeared in were accessed through the official COCA website. The COCA website allows a search function that shows every occurrence of the idiom in the corpus along with the surrounding contexts. This judgement was made by a non-native English speaker; by inferring the intended meaning of the idioms using the surrounding context. The final frequency of the idiom was solely based on the number of figurative use instead of the raw frequency, because the present study is interested in how frequently idioms are used figuratively in children's language.

3 Results

In pursuit of the research question: how frequently are idiomatic expressions present in English juvenile fiction, the present study found 46 idiomatic N-grams that occurred at least 5 times in the corpus. The idiomatic N-grams were found amongst the combination of 11.708 unique N-grams from the original text and 16.187 unique N-grams from the lemmatized text.

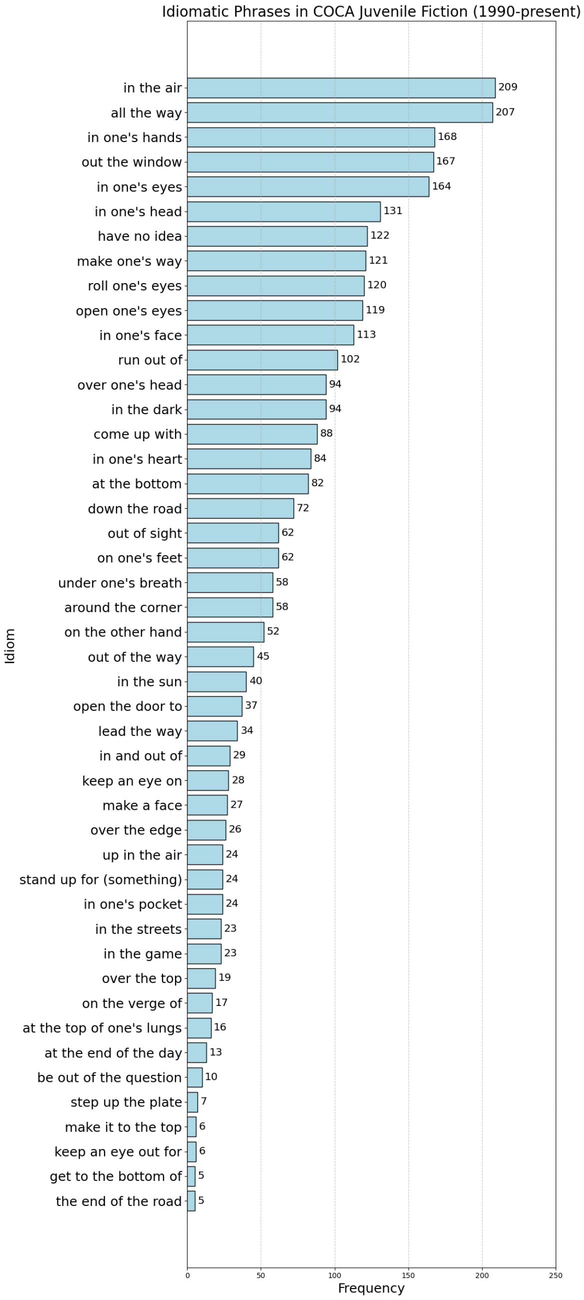

Figure 3.1: Raw frequency distribution of the idioms

The full list of 46 idioms can be seen in the Appendix; along with the figurative frequency counts, appearance counts in books and magazines, and the definitions of the idioms. The idioms identified are exclusively 3-grams, 4-grams, 5-grams, and 6-grams; indicating that no idioms were detected among N-grams of degrees 7 to 10. The most frequent idiom, namely *in the air*, is a 3-gram and occurred 209 times in the corpus; while the least frequent idiom, *the end of the road*, is a 5-gram and occurred 5 times (which may be due to the *min_occur* of 5 for 5-grams). The lowest *min_occur* value was 3 for 6- to 10-grams. Although 6-gram idioms were found, their occurrences were more than 3 times.

Figure 3.1 shows the top most frequent idioms ranked based on their respective counts. The frequency at which each idiom was used figuratively is illustrated in Figure 3.2. It should be noted that the order of the idioms on the y-axis of the plot is now different compared to Figure 3.1.

The most frequent, figuratively used idiom *have no idea* was used 122 times in a figurative context. In contrast, the idioms *at the end of the day, lead the way, in the sun* and *in one's pocket* were used figuratively the least, with 0 figurative uses.

By the definition of an idiom in this study, an idiom may have both a literal and a non-literal meaning. The use of an idiom may then also be non-literal or literal. To verify that the zero figuratively used idioms do indeed have idiomatic uses, these idioms were explored in the general COCA corpus and analyzed whether they were used idiomatically in adult language.

The fiction section of COCA was selected for the aforementioned analysis and 100 random samples for each of the idioms *at the end of the day, lead the way, in the sun* and *in one's pocket* were analyzed. The figurative frequency counts of these idioms can be observed in Table 3.1 Based on what can be seen in Table 3.1, the idioms that were not used figuratively at all in the juvenile fiction corpus, indeed have figurative uses in the general fiction corpus.

A comparison between the figurative and literal use of the idioms is presented in Figure 3.3. The distribution of the raw frequency is as expected, where certain idioms occur more frequently than others. The proportion of figurative and raw frequency vary across all idioms; certain idioms were more likely to be used literally than figuratively,

Idiom	Figurative use out of 100 random samples	Total occurences
At the end of the day	30	440
Lead the way	11	165
In the sun	1	1889
In one's pocket	5	3222

Table 3.1: Figurative use of the idioms in the fiction genre of the entire COCA corpus. These idioms were only used literally in the juvenile fiction corpus.

while certain idioms were used mostly figuratively. Overall, the figuratively used idioms appeared less in the corpus.

There seems to be no apparent pattern that is causing the differences in the proportion of raw and figurative frequency across the idioms. To see if there are differences caused by the characteristics of the idioms themselves, the top 5, middle 5, and bottom 5 idioms were listed based on figurative frequency in Table 3.2. As suggested by (Nordmann et al., 2014), decomposability ratings can vary between individuals and therefore is not a reliable measure of an idiom's decomposability. With this limit in mind, the current study can only suggest judgements on the idioms decomposability based on the author's, non-native English knowledge (which may bear differences with the judgement of a native speaker).

From Table 3.2, it appears that majority of the idioms are decomposable; meaning the figurative meaning of the idioms are closely related to the literal meaning of the composing words (such as *have no idea, in one's hand, out of the way, at the end of the day, etc.*). However, the idioms *roll one's eyes, be out of the question* and *in the sun*, are less decomposable. Both less decomposable and more decomposable idioms are spread across the top, middle, and bottom idioms; therefore, there appears to be no pattern in the figurative frequency differences caused by decomposability.

Top 5 idioms	Middle 5 idioms	Bottom 5 idioms
have no idea	out of the way	the end of the road
roll one's eyes	in and out of	in one's pocket
come up with	be out of the question	at the end of the day
in one's hand	make one's way	in the sun
in one's heart	in the dark	lead the way

Table 3.2: Comparison of the top, middle, and bottom 5 idioms based on figurative frequency

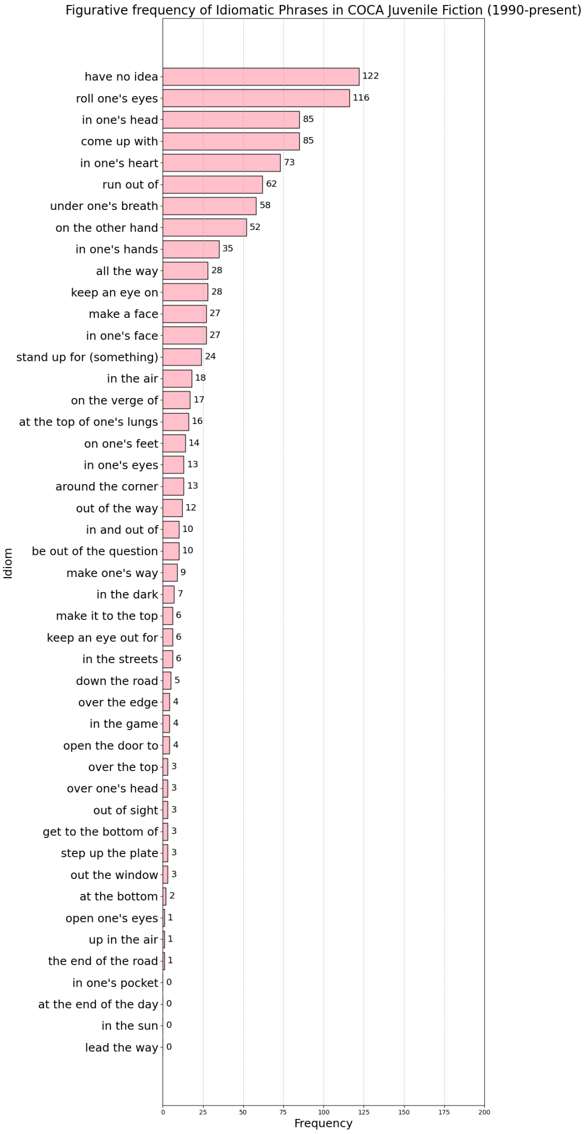

Figure 3.2: Figurative frequency distribution of the figuratively used idioms

4 Discussion

To summarize, the current study aimed to find frequent idioms from a juvenile fiction corpus, with the purpose of investigating how frequently present idioms are in children's literature. Thus, this study explored a different method of operationalizing idiom exposure in children and the implications on children's idiom exposure through this particular linguistic environment. From this pursuit, a total of

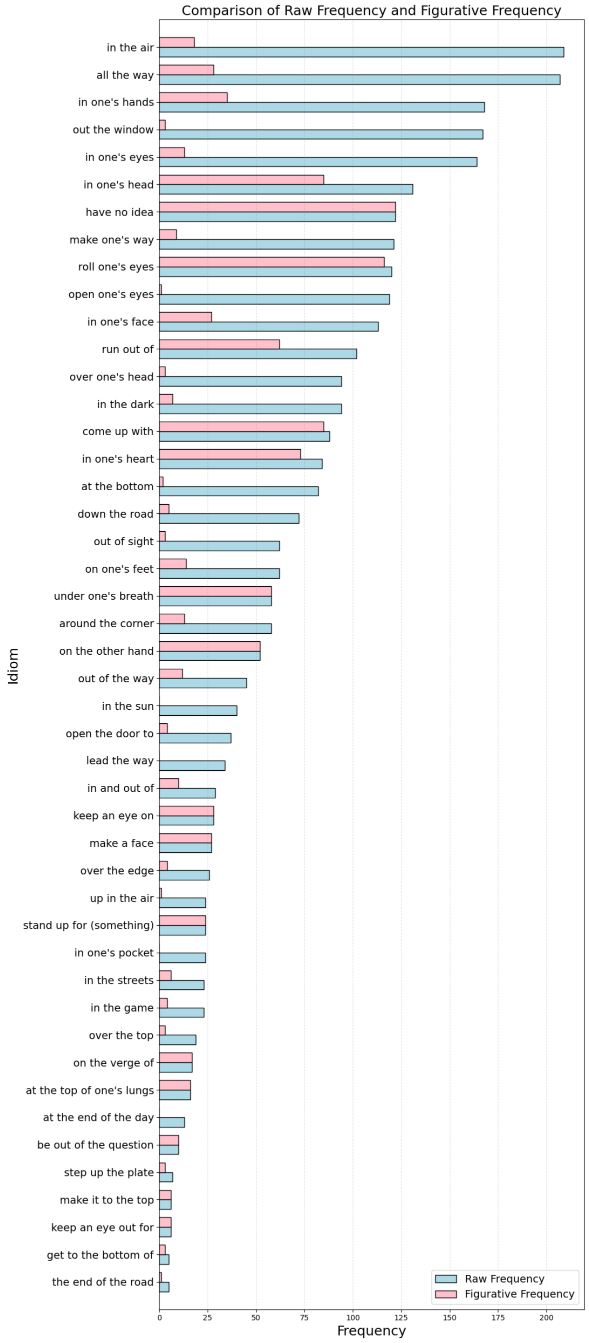

Figure 3.3: Comparison of the raw frequency and figurative frequency of the idioms, ordered based on raw frequency count

46 idioms were found from the 3 million word, juvenile fiction corpus. The idioms spanned between 3- to 6-grams, with all idioms occurring at least 5 times and the most frequent idiom occurring 209 times. The overall raw frequencies of the idioms are higher than the figurative frequencies, where the highest figurative frequency is 122 occurrences.

To answer the research question: how frequently are idiomatic expressions present in English Juvenile fiction; the findings of the current study suggests that the frequency of idioms in American English juvenile fiction vary between idioms. The distribution of both the raw and figurative frequency shown in the results highlight that certain idioms are highly frequent, thus more often used compared to the other retrieved idioms.

The results contained more idioms that were seldom used figuratively in proportion to their raw frequencies. This finding could suggest that the idioms children are exposed to in juvenile fiction are more likely to be used literally. The creators of these fictions may have chosen to not use idioms in a figurative sense to account for readers that may not understand idioms figuratively. The current study could not access the different target age ranges of the fictional works, hence any possible relation between the target reader's age and the creator's choice of idioms could not be explored.

Explored previously in the results section: decomposability (subjective to this study) of the idioms appeared to not differ across the idioms, despite their figurative frequency counts. This may suggest that decomposability of idioms does not affect the figurative use of idioms in children's literature. Since this conclusion is drawn from a subjective measure, it is recommended for further studies to investigate the role of decomposability in the idioms figurative frequency; by combining ratings from more than one individual.

Referring back to the findings of research in children's idiom knowledge development, the general observation is that children of under the juvenile age category (7 to 18-years old) have sufficient abilities to understand certain idioms. This applies whether one believes in the mere exposure effect or the figurative competence viewpoint. However, from the results of the current study, literal uses of idioms appear more out of all idiom uses. This raises the question of how do children then encounter and acquire figuratively used idioms, if idioms are used more literally in juvenile fiction?

If the aforementioned finding applies beyond the sample of juvenile fiction used in this study, it may imply that children learn the figurative meanings of idioms through a different source. Possible sources of this information is through other linguistic environments (education or other non-fiction sources, movies, TV shows, etc.) or through spoken language. Further endeavors to answer this question should be made to draw reliable conclusions on how and where children are more exposed to figuratively used idioms. By answering this question, the operationalization of exposure can be extended reliably as an addition to familiarity ratings.

Overall, the findings mentioned would imply that children are seldom exposed to idioms in a non-literal way through written fiction. Exposure to more literal uses of idioms may influence the way children learn the figurative meaning of idioms. Hence, the creator's choice of idiom usage in juvenile fiction can contribute to the understanding of how exposure to idioms affect children's idiom knowledge. In addition to familiarity ratings, taking into account the possible exposure children receive from written form of fiction may add valuable insight in measuring children's idiom exposure.

Through the current study, analyzing idiom frequencies as a proxy to children's idiom exposure have shown that a bottom-up approach is beneficial in uncovering how idioms are present in children's linguistic environment. This method shows that idioms appear and may be exposed to children not only figuratively, but also literally. An idiom may then be familiar to a child, but it does not warrant familiarity to the figurative meaning of the idiom. Hence, exploring idiom presence through a bottom-up approach can give insight into how researchers should select idioms for familiarity ratings.

Despite the potentially insightful findings, the current study bears limitations. The current study took a bottom-up approach in investigating idiom frequency in the corpus, which was done by first extracting phrases (in the form of N-grams) and then analyzing the idiomaticity of the phrases. This approach was chosen as it allows any idiomatic expression to surface from the data, as opposed to searching for a set of predetermined idioms. The idioms that surfaced due to this method are idioms that were naturally frequent in the corpus, which suitably answers the question of how frequently idioms are present in juvenile fiction.

Albeit the advantages, a bottom-up approach faces several limitations as well. The frequency threshold chosen for each N-gram determines which N-grams, and therefore which idioms, are discovered from the data. In the current study, this frequency threshold was set by the *min_occur* parameter (minimum frequency). As a reminder, the 3- and 4-grams had *min_occur* values of 15 and 14, respectively; which is a higher *min_occur* value compared to the other N-grams with *min_occur* values of 3 (6- to 10-grams) or 5 (5-grams). The idioms found occurred at least 5 times, which means no idioms were found from N-grams that occurred at least 3 or 4 times in the corpus. Hence, it is likely that no idioms were overlooked from N-grams with a *min_occur* of 3. However, the higher frequency threshold in 3- to 5-grams might have resulted in potential idioms being missed.

Nevertheless, the frequency threshold was necessary to implement in order to control the number of N-grams retrieved from the data. In the case of 3- and 4-grams for example, a lower *min_occur* value would have resulted in thousands of N-grams to be inspected, which is beyond the scope of this project. The *min_occur* was therefore set to an optimal value that would generate N-grams with a low minimum frequency while still keeping the number of N-grams feasible for manual inspection. As a consequence, the idioms found in this study are only a portion of the idioms that could be found in the data, especially from the N-grams that had a high *min_occur* value like 3- to 5-grams.

In addition, the current approach did not explore N-grams beyond length 10. Although idioms were not discovered from N-grams longer than 6-grams; there still lies a possibility of finding idioms from N-grams longer than the lengths explored in the current study. Due to constraints in the current project, exploring higher degree N-grams or a larger number of N-grams with lower minimum frequencies remain for future projects to undertake.

To judge the idiomaticity of the retrieved N-grams, the lists of N-grams had to be manually inspected. In this process, human error is a risk factor that is challenging to detect. Due to the extensiveness of the list, traversing through each N-gram is a time consuming process that requires the inspector to maintain consistent attention. Future projects that aim to replicate the current project would benefit from employing more than one inspector to reduce human error. Additionally, inspecting the N-gram lists more than once would reduce the risk of overlooking any idioms and result in more precise findings. In this study, each list of N-grams (both from the original and lemmatized data) were inspected twice to reduce this error, although only done by one individual.

Lastly, the idioms found in this study only reflect what idioms are used in the language and genres of fiction covered in the corpus. The juvenile fiction corpus consists of data written in American English and covers fiction in the form of books and magazines. Other forms of English, such as British English, may have different idioms due to cultural and societal differences within the British English speakers. This may also have affected the idiomaticity judgement process as it was partially based on a British English speaker. Therefore, utilizing a corpus with a different variation in English and relying on different types of English speakers could result in different findings.

Furthermore, the corpus data only covers fiction in written form. Spoken language is also an interesting area to discover idioms from and can reveal how frequently idioms appear in speech used around children. Spoken language may also show a different proportion between the figurative and raw frequencies of the idioms compared to the proportions found in this study.

In summary, the frequency at which idioms are present in juvenile fiction varies across all idioms. Certain idioms are more frequently used than others, which is commonly the case for words and phrases in natural language; not all words, phrases, or idioms are used at the same frequency. Although, the most frequent idioms were not necessarily used figuratively. The current study could not explore factors that might have caused the varying proportions of figurative and literal uses of the idioms, except for a subjective decomposability rating. This decomposability rating also did not imply any effects on the varying frequency proportions.

Further studies on more extensive N-gram lists, on different types of fiction, or on spoken language data would contribute more in uncovering the degree of idiom exposure children receive from their environment. Attempting to operationalize idiom exposure in children would consequently advance the understanding of how exposure plays a role in children's idiom knowledge. Analysis on the characteristics of the idioms that surface may also shed light to which and why certain idioms are used more figuratively than others in language used towards children.

References

Abkarian, G. G., Jones, A., & West, G. (1992, June). Young Children's Idiom Comprehension: Trying To Get the Picture. *Journal of Speech, Language, and Hearing Research*, *35*(3), 580–587. Retrieved 2023-11-21, from `http://pubs.asha.org/doi/10.1044/jshr.3503.580` doi: 10.1044/jshr.3503.580

Ammer, C. (2013). *The American Heritage dictionary of idioms* (Second edition ed.). Boston: Houghton Mifflin Harcourt.

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based approaches to issues in applied linguistics. *Applied linguistics*, *15*(2), 169–189.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Chafe, W. L. (1970). *Meaning and the structure of language.* Chicago: University of Chicago Press.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA).* Retrieved 2024-01-17, from `https://www.english-corpora.org/coca/`

Ezell, H. K., & Goldstein, H. (1991). Comparison of idiom comprehension of normal children and children with mental retardation. *Journal of Speech, Language, and Hearing Research*, *34*(4), 812–819.

Farlex. (2024). *The Free Dictionary.* Retrieved 2024-01-17, from `https://idioms.thefreedictionary.com/`

Gibbs, R. W. (1987, October). Linguistic factors in children's understanding of idioms. *Journal of Child Language*, *14*(3), 569–586. Retrieved 2023-11-21, from `https://www.cambridge.org/core/product/identifier/S0305000900010291/type/journal_article` doi: 10.1017/S0305000900010291

Lazar, R. T., Warr-Leeper, G. A., Nicholson, C. B., & Johnson, S. (1989, October). Elementary School Teachers' Use of Multiple Meaning Expressions. *Language, Speech, and Hearing Services in Schools*, *20*(4), 420–430. Retrieved 2023-11-29, from http://pubs.asha.org/doi/10.1044/0161-1461.2004.420 doi: 10.1044/0161-1461.2004.420

Levorato, M. C., & Cacciari, C. (1992). Children's comprehension and production of idioms: the role of context and familiarity. *Journal of Child Language*, *19*(2), 415–433.

Levorato, M. C., & Cacciari, C. (1995). The effects of different tasks on the comprehension and production of idioms in children. *Journal of experimental child psychology*, *60*(2), 261–283.

Libben, M. R., & Titone, D. A. (2008). The multi-determined nature of idiom processing. *Memory & cognition*, *36*, 1103–1121.

Liu, D. (2003). The most frequently used spoken american english idioms: A corpus analysis and its implications. *Tesol Quarterly*, *37*(4), 671–700.

Lodge, D. N., & Leach, E. A. (1975, September). Children's Acquisition of Idioms in the English Language. *Journal of Speech and Hearing Research*, *18*(3), 521–529. Retrieved 2023-11-29, from http://pubs.asha.org/doi/10.1044/jshr.1803.521 doi: 10.1044/jshr.1803.521

Martinez, R., & Schmitt, N. (2012, July). A Phrasal Expressions List. *Applied Linguistics*, *33*(3), 299–320. Retrieved 2023-11-27, from https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/ams010 doi: 10.1093/applin/ams010

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching.* Oxford University Press.

Nippold, M. (1991, July). Evaluating and Enhancing Idiom Comprehension in Language-Disordered Students. *Language, Speech, and Hearing Services in Schools*, *22*(3), 100–106. Retrieved 2023-11-29, from http://pubs.asha.org/doi/10.1044/0161-1461.2203.100 doi: 10.1044/0161-1461.2203.100

Nippold, M. (2006, 12). Language development in school-age children, adolescents, and adults. *Encyclopedia of Language Linguistics*. doi: 10.1016/B0-08-044854-2/00852-X

Nippold, M., & Martin, S. T. (1989). Idiom interpretation in isolation versus context: A developmental study with adolescents. *Journal of Speech, Language, and Hearing Research*, *32*(1), 59–66.

Nippold, M., & Taylor, C. L. (2002, April). Judgments of Idiom Familiarity and Transparency: A Comparison of Children and Adolescents. *Journal of Speech, Language, and Hearing Research*, *45*(2), 384–391. Retrieved 2023-11-21, from http://pubs.asha.org/doi/10.1044/1092-4388%282002/030%29 doi: 10.1044/1092-4388(2002/030)

Nordmann, E., Cleland, A. A., & Bull, R. (2014). Familiarity breeds dissent: Reliability analyses for british-english idioms on measures of familiarity, meaning, literality, and decomposability. *Acta psychologica*, *149*, 87–95.

Pawley, A., & Syder, F. (1983, 01). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication.*

Spector, C. C. (1996, October). Children's Comprehension of Idioms in the Context of Humor. *Language, Speech, and Hearing Services in Schools*, *27*(4), 307–313. Retrieved 2023-11-21, from http://pubs.asha.org/doi/10.1044/0161-1461.2704.307 doi: 10.1044/0161-1461.2704.307

Sprenger, S. A., la Roi, A., & van Rij, J. (2019). The Development of Idiom Knowledge Across the Lifespan. *Frontiers in Communication*, *4*. Retrieved 2023-12-04, from https://www.frontiersin.org/articles/10.3389/fcomm.2019.00029

Tabossi, P., Arduino, L., & Fanari, R. (2011). Descriptive norms for 245 italian idiomatic expressions. *Behavior Research Methods*, *43*, 110–123.

Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol*, *9*(4), 247–270.

van Rij, J., Uithof, F., Poelstra, S., M. Jones, S., & Sprenger, S. (2023). Adding a piece to the puzzle: Children's exposure to idioms. *Unpublished manuscript*.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual.* Scotts Valley, CA: CreateSpace.

A Appendix

Idiom	Figurative frequency	Book	Magazine
in the air	18	11	7
all the way	28	10	18
in * hands	35	18	17
out the window	3	3	0
in * eyes	13	9	4
in * head	85	54	31
have no idea	122	54	68
make one's way	9	8	1
roll * eyes	116	51	65
open * eyes	1	0	1
in one's face	27	14	13
run out of	62	11	51
over * head	3	2	1
in the dark	7	4	3
come up with	85	31	54
in one's heart	73	46	26
at the bottom	2	0	2
down the road	5	2	3
out of sight	3	1	2
on one's feet	14	5	9
around the corner	13	10	3
under * breath	58	43	15
on the other hand	52	40	12
out of the way	12	6	6
in the sun	0	0	0
open the door to	4	2	2
lead the way	0	0	0
in and out of	10	4	6
keep an eye on	28	16	12
make a face	27	17	10
over the edge	4	3	1
in one's pocket	0	0	0
stand up for *	24	1	23
up in the air	1	1	0
in the streets	6	4	2
in the game	4	2	2
over the top	3	3	0
on the verge of	17	11	6
at the top of * lungs	16	4	12
at the end of the day	0	0	0
be out of the question	10	5	5
step up the plate	3	0	3
keep an eye out for	6	1	5
make it to the top	6	3	3
get to the bottom of	3	2	1
the end of the road	1	1	0
Total:	1019	513	505

Table A.1: Figurative frequency distribution of each idiom across juvenile fiction types. The asterisk denotes a variable component of the phrase

Idiom	Definition
in the air	In circulation, in people's thoughts
all the way	Completely, from start to finish
in one's hands	In one's responsibility, charge, or care
out the window	Forgotten, disregarded, wasted
in one's eyes	In someone's estimation or opinion
in one's head	Not substantial or real; having been imagined or invented
have no idea	To be completely unaware of or know no information about something
make one's way	Go in a particular direction or to a particular destination
roll one's eyes	To turn one's eyes upward or around in a circle, usually as an expression of exasperation, annoyance, or impatience
open one's eyes	Become or make someone aware of the truth of a situation
in one's face	Defiantly confrontational; also, an exclamation of contempt
run out of	Exhaust a supply or quantity of
over one's head	To a position higher than another's; 2. Beyond one's understanding or competence
in the dark	In secret, in concealment; 2. In a state of ignorance, uninformed
come up with	Produce, supply, discover
in one's heart	Produce, supply
at the bottom	To discover the origin of a problem, or the fundamental truth of an issue or event
down the road	Ultimately; when everything else has been taken into consideration
out of sight	Unreasonable, excessive; 2. Excellent, superb
on one's feet	In a healthy or stable condition, usually after a period of illness or misfortune; 2. Quickly or extemporaneously
around the corner	Nearby, a short distance away 2. Very soon, imminent
under one's breath	Softly, in an undertone or whisper
on the other hand	From a different, conflicting, or contradictory point of view
out of the way	Not obstructing, hindering, or interfering; 2. Taken care of, disposed of; 3. In a remote location
in the sun	Receiving the public's scrutiny or attention
open the door to	Create an opportunity for
lead the way	Act as a guide; 2. Be first or most prominent in some field or action
in and out of	Being a frequent participant in a certain situation or place
keep an eye on	To watch over attentively; mind 2. To watch closely or carefully
make a face	Grimace, distort the facial features
over the edge	Beyond a certain limit, threshold, goal, or quota; 2. Into a state of emotional instability
in one's pocket	In one's power or possession, under one's influence
stand up for (someone or something)	Side with, defend
up in the air	Not settled, uncertain
in the streets	Without an established place of residence or accommodation; homeless; 2. In a state of being widely and publicly known or discussed
in the game	Actively participating in something
over the top	Surpassing a goal or quota 2. Extreme, outrageous
on the verge of	Close to, on the brink of
at the top of one's lungs	With an extremely loud voice
at the end of the day	Ultimately, in the end
be out of the question	To be impossible and/or impermissible
step up the plate	Take action in response to an opportunity or crisis
keep an eye out for	To remain vigilant or carefully watchful for something or someone
make it to the top	To win
get to the bottom of	To determine the cause or source of something or solve the mystery of something
the end of the road	The conclusion or final step of something

Table A.2: Definitions of the 46 idioms