FAKULTA
PŘÍRODOVĚDNĚ-HUMANITNÍ
A PEDAGOGICKÁ TUL

Bakalářská práce

Frequent words and keywords in Terry Pratchett's novel Colour of Magic

Studijní program: B0114A300068 Anglický jazyk se zaměřením na vzdělávání
Studijní obory: Anglický jazyk se zaměřením na vzdělávání
Španělský jazyk se zaměřením na vzdělávání

Autor práce: **Vojtěch Hrůza**
Vedoucí práce: Mgr. Petra Peldová, Ph.D.
Katedra anglického jazyka

Liberec 2023

FAKULTA
PŘÍRODOVĚDNĚ-HUMANITNÍ
A PEDAGOGICKÁ TUL

Zadání bakalářské práce

Frequent words and keywords in Terry Pratchett's novel Colour of Magic

Jméno a příjmení:	**Vojtěch Hrůza**
Osobní číslo:	P19000191
Studijní program:	B0114A300068 Anglický jazyk se zaměřením na vzdělávání
Specializace:	Anglický jazyk se zaměřením na vzdělávání Španělský jazyk se zaměřením na vzdělávání
Zadávající katedra:	Katedra anglického jazyka
Akademický rok:	2020/2021

Zásady pro vypracování:

Obsahem bakalářské práce je korpusově stylistická analýza knihy "The Colour of Magic" od Terryho Pratchetta. V práci bude uplatněn jak kvantitativní tak kvalitativní přístup. Analýza bude postavena na seznamech frekventovaných a klíčových slov a jejich následné detailní analýzy z pohledu kolokací a n-gramů. Autor se také zaměří na lexikální hustotu (Ld) románu, kdy Ld poměřuje obsahové a gramatické složky. Výsledkem analýzy bude popsání idiolektu Terryho Pracheta v daném románu. Jako referenční korpus bude sloužit korpus složený z knih různých autorů stejného žánru.

Rozsah grafických prací:
Rozsah pracovní zprávy:
Forma zpracování práce: tištěná/elektronická
Jazyk práce: Angličtina

Seznam odborné literatury:

- Pratchett, Terry, "The Colour of Magic." London: Transworld Publishers, 1985
- Luthi, Daniel. "Toying with fantasy: the postmodern playground of Terry Pratchett's Discworld novels." Mythlore: A Journal of JRR Tolkien, CS Lewis, Charles Williams, and Mythopoeic Literature 33, no. 1 (2014): 8. https://dc.swosu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1036&cont
- Ukrayinska, Y. V. "Linguistic techniques of creating humour of magic in T. Pratchett's novels." Science and Education a New Dimension. Philology, VIII 69, no. 235 (2020): 54-59. http://seanewdim.com/uploads/3/4/5/1/34511564/httpsdoi.org10.31174send-ph2020--235viii70-13.pdf
- Oxford University Press, "Oxford Advanced Learner's Dictionary." accessed June 6, 2021. http://www.oxfordadvancedlearnersdictionary.com
- Bowker, Lynne. "Corpus Linguistics is Not just for Linguists: Considering the Potential of Computer-Based Corpus Methods for Library and Information Science Research." Library Hi Tech 36, no. 2 (2018): 358-371, https://www.proquest.com/scholarly-journals/corpus--linguistics-is-not-just-linguists/docview/2036138529/se-2?accountid=17116 (accessed June 6, 2021).
- Kilgarriff, Adam. "Simple maths for keywords." In Proceedings of the Corpus Linguistics Conference. Liverpool, UK. 2009. https://www.sketchengine.eu/wp--content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf
- Kilgarriff, Adam, and Pavel Rychlý. "Sketch Engine User Guide." Sketch Engine. accessed June 6, 2021. https://www.sketchengine.eu/guide/

Vedoucí práce: Mgr. Petra Peldová, Ph.D.
Katedra anglického jazyka

Datum zadání práce: 19. června 2021
Předpokládaný termín odevzdání: 15. července 2022

L.S.

prof. RNDr. Jan Picek, CSc.
děkan

Mgr. Zénó Vernyik, Ph.D.
vedoucí katedry

V Liberci dne 19. června 2021

Prohlášení

Prohlašuji, že svou bakalářskou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Jsem si vědom toho, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má bakalářská práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

25. dubna 2023 Vojtěch Hrůza

Poděkování

Chtěl bych poděkovat své vedoucí bakalářské práce Mgr. Petře Peldové, Ph.D. za odborné vedení, pomoc, ochotu a rady při zpracování této práce.

Abstract

The content of this bachelor thesis is a corpus analysis of the book "The Colour of Magic" by Terry Pratchett. Firstly, the thesis describes the basic terms and methods associated with corpus linguistics, which are used in the analysis. Subsequently, hypotheses are formed from the obtained data, concerning the context of the book, the choice of words, the frequent occurrences of certain phrases, etc. The hypotheses are then examined and either confirmed or disproved. The main objective of this work should be to obtain new information about the language structure of the work and to reveal particular tendencies and habits the author resorted to when writing the book.

Obsahem bakalářské práce je korpusově stylistická analýza knihy “The Colour of Magic“ od Terryho Pratchetta. V práci bude uplatněn jak kvantitativní tak kvalitativní přístup. Analýza bude postavena na seznamech frekventovaných a klíčových slov a jejich následné detailní analýzy z pohledu kolokací a n-gramů. Autor se také zaměří na lexikální hustotu (Ld) románu, kdy Ld poměřuje obsahové a gramatické složky. Výsledkem analýzy bude popsání idiolektu Terryho Pracheta v daném románu. Jako referenční korpus bude sloužit korpus složený z knih různých autorů stejného žánru.

Contents

Figures

Tables

Abbreviations

BNC – British National Corpus

COCA – Corpus of Contemporary American English

LOB – Lancaster-Oslo-Bergen Corpus

TTR – Type-token ratio

1. Introduction

This paper describes quantitative and qualitative research investigating the linguistic entirety of the book Colour of Magic by Terry Pratchett, published in 1983. The practical part of the thesis consists of two sections. In the first one, the objective is to uncover specific features of the author's idiolect, detecting whether there are any observable patterns of his style or if the choice of words is specific for any of the characters, settings, or even the story as a whole. It is presumed that Terry Pratchett, being such a worldwide acknowledged writer, must have created a personal style to get the readers´ attention. In addition, the first section investigates the most frequently used lemmas, their distribution among the text, and their most frequent collocates. Attention is given to N-grams, which should provide a further insight into Pratchett's preferred vocabulary. The second part then compares the described idiolect with other books of the same genre (fantasy), which were published between 1980 and 1989. Here, the main objective is to determine whether or not the discovered features of Pratchett's writing style are unique or if he shares them with other writers of a similar time period and genre.

Since it could be considered obvious that no two idiolects are completely identical, the goal is to specifically describe in which ways Pratchett differs from his peers. The results could then be used further by other academics and their research involving either the same book, author or time period for the particular genre. Therefore, a tertiary objective of this paper is to provide information which could then be used as a topic of another thesis, investigating the phenomena in greater detail or in a different context. The theoretical part of the thesis synthesises all the theoretical background of the research, namely corpus linguistics and its aspects, such as n-grams, wordlists, keywords, and lexical density.

1.1. Idiolect

According to Oxford dictionary, an idiolect is "the way that a particular person uses language." Such a definition, however, might be a bit too straightforward. Louwerse (2004: 207) offers a more detailed explanation: "Idiolects are person-dependent similarities in language use. They imply that texts by one author show more similarities in language use than texts between authors." Therefore, when examining a text by a single author and comparing it to texts by other authors, one should notice a series of habits and patterns within the chosen author's work. As described in chapter 1, this paper partially focuses on such comparison.

1.1.1. Pratchett's idiolect

Since one of the aforementioned objectives of this paper is to attempt to identify Pratchett's idiolect based on the data gathered from the corpora, the first step was to investigate how other academics have described the author's idiolect in the past.

Arguably the most prominent feature of Pratchett's writing is the element of humour, which is often added via the means of parody and satire. Parody can be mostly seen through characters, which are often described as complete opposites of the tropes they are meant to represent. An example of this is the wizard Rincewind, who is described as someone lacking in intelligence, contradictory to the wise wizards seen in other literature of the same genre (Broeder, 2007). Satire is then present mainly in plot and world building, being used to criticize certain aspects of the real world that the author himself had negative experiences with (Britton, 2018). For example, his time spent working as a journalist most likely contributed to the following sentence: "[The city] was […] denied the benefit of newspapers, leaving the population to fool themselves as best they could" (Pratchett, 1996: 145).

These elements of Pratchett's idiolect are closely connected to larger fragments of text. This paper, however, mainly deals with much smaller units, such as individual tokens, their collocations, N-grams, etc.

The first research question is therefore following: Can the aforementioned elements of Pratchett's writing, i.e., parody and satire, be somehow observed through the data obtained from this work's focus corpus?

Other sub questions are consequently stated in the practical part through the analysis.

2. Theoretical part

2.1. Corpus linguistics

The first necessary step to describing the theory behind this paper is to explain the basic features of corpus linguistics as stated above. The name itself already provides a certain basis for such an explanation. "Linguistics is the scientific study of language [and] a corpus is a large collection of naturally occurring texts (e.g., those which were not originally created for the purposes of language analysis)" (Baker and Egbert, 2020: 3). From this quote, it is therefore apparent that this particular approach of linguistics focuses on work with large number of computerised texts. In order to imagine the potential size of a corpus, a few examples of corpora

are now presented. The first officially recognised representative corpus of the American English, the Brown corpus, and its counterpart representing the British English, the LOB corpus, both consisting of 1 million tokens were viewed as "huge" at the time of their release (1960s). As the time passed the British National Corpus (BNC) was compiled and released in the late 1990s, it comprises 100 million tokens (Biber and Reppen 2015: 11). The sizes of corpora have been increasing due to the fast development of digital technologies. For example, the Corpus of Contemporary American English (COCA) has 450 million tokens, and it is not considered huge today. The Sketch Engine platform, which is used for the purposes of this paper, currently offers more than 710 different language corpora ranging from 13,000 tokens to more than 60 billion tokens.

Such impressive amounts of tokens are processed digitally; they are annotated with the help of very accurate, efficient, and reliable up-to-date software which is based on modern computational technologies (Baker and Egbert, 2020). Detailed linguistic research would be impossible without such technologies (Lindquist, 2009). The computer, therefore, works as a tool which gathers the data and directs the researcher, for further manual analysis, to certain aspects of the corpus. It allows the storage of large volumes of texts, which can then be quickly searched through for particular instances of words, phrases, and patterns. These can then be sorted, filtered, and displayed for the researcher, thus making working with the collected data more straightforward (Bowker, 2018). Yet, it must be emphasised that human "digging deeper" cannot be omitted. The technologies are here to help the linguist, but they will not replace them.

The above-mentioned corpora were used as examples of representative corpora; however, corpora can generally differ in nature. They can be not only representative, but also specialised, historical, academic etc. These are used if one wishes to restrict their results to a specific genre or register. An example of both specialised and academic corpora is the MICASE (Michigan Corpus of Academic Spoken English) corpus, which was created out of recorded and transcribed spoken English samples collected at the University of Michigan (contains about 1.8 million words). A significant historical corpus is then, for example, the Lampeter Corpus of Early Modern English tracts, containing 120 texts from tracts published between 1640 and 1740 (Lindquist, 2009).

Furthermore, corpora can also be divided into two categories based on their form of language, i.e., spoken and written. In the case of a spoken corpus, the samples are gathered via recorded media or recording devices recorders and then transcribed with the assistance of a software. Current speech recognition technologies are, however, not yet sufficiently developed to

independently create error-free transcriptions, so they have to be accompanied by significant manual checking (Biber and Reppen, 2015). Due to the complexity of such a method, spoken corpora are lagging behind the written ones, which can be created rather easily from texts available online (Lindquist, 2009), or formed out of scanned hardcopies, having been edited by software to create a machine-readable version of the text from provided image. Even in this case, however, the software is not completely reliable, and the scanned text has to be edited and adjusted to avoid too many errors. In certain cases, even such measures might not be sufficient and mistakes may still appear in the resulting files. Therefore, just as in the case of the spoken corpora, manual checking is necessary (Biber and Reppen, 2015).

2.2. Types, tokens, lemmas, and word lists

When corpora are analysed in details, their wordlists are created because they offer insight into the most frequently used words in a corpus. However, using the term *word* can be rather misleading. As Brezina (2018: 39) points out:

> "In order to count words reliably we, of course, need to know what we are counting, and therefore we first have to define a word. This might look fairly trivial (we all know what a word is), but in fact the definition of a word is quite a complex issue."

One example of the problems that may be encountered is the question of dealing with compounds. How would the words *blow torch*, *blowtorch,* and *blow-torch* be counted? There is no concrete answer to this question as it all depends on the definition (Lindquist, 2009: 35). Therefore, linguists use three different terms to describe what exactly is counted when working with wordlists and their content frequencies. These terms are tokens, types, and lemmas (Brezina, 2018). Token is a "string of letters or numbers separated by a white space (or punctuation)" (ibid.: 39). Figure 1 serves as an example. Here, it is possible to see how tokens are counted, resulting in a total of 26.

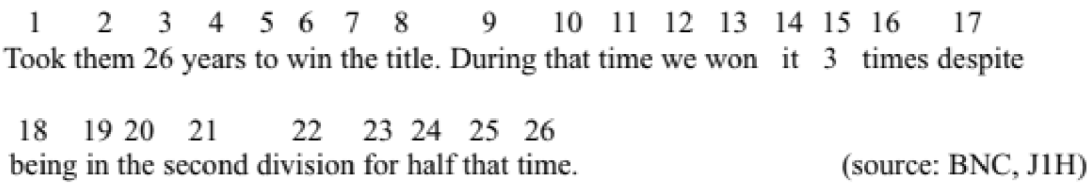

Figure 1: Examples of counting tokens

Types, on the other hand, describe how many different words are in a corpus. This means that two or more instances of the same token are counted as only one type (Lindquist, 2009).

Returning to figure 2 we can see that tokens *the* (7, 20), *that* (10, 25) and *time* (11, 26) each occur twice, and therefore, the total number of types in this example is 23 (Brezina, 2018). A different category altogether are lemmas. Lemmatisation is a process in which various inflections of a word are grouped together and classified as one word (Lindquist, 2009). What is counted are therefore not the individual tokens, but rather their basic forms, which correspond to what one might find in a dictionary, see example (1) (Kilgarriff and Rychlý, 2021).

(1) Lemma *go* = types *go*, *went*, *gone*, *goes*, *going*

When working with word lists, a linguist can then decide if they wish to see lists of types tokens or lemmas, depending on their research focus. Apart from these aspects, it is also important to be aware of the difference between grammatical and content/lexical words. Grammatical words are

> "words for which the primary function is to indicate grammatical relationships, as distinct from lexical words, the primary function of which is referential (content words). Grammatical words include, [for example], articles, pronouns, and conjunctions. Lexical words include, [for example], nouns, verbs, and adjectives." (Chandler and Munday, 2011: 178)

Due to the difference between these two types of expressions, assumptions can be formed regarding the likely contents of a word list. Grammatical words are often repeated due to their natural usage in English language. Usually, it is not possible to form a sentence without them. Therefore, a rather large presence of grammatical words in a word list is to be expected (Brezina, 2018).

2.3. Keywords

"Keywords are words which are statistically significantly more frequent in one corpus (or subset of a corpus) when compared against another corpus (or subset thereof)" (Baker and Egbert, 2016: 12). Therefore, in a keyword approach to a corpus analysis, the researcher works with two corpora. One functions as the primary subject of the study, while the other, the reference corpus, is used as a kind of filter. Frequencies of all lexical expressions from both corpora are compared and those that occur almost the same number of times are then filtered out, as they are not present significantly more frequently in only one corpus (Scott and Tribble, 2006). What is therefore displayed in a keyword list is a collection of words, that are used nearly exclusively in either the focus or the reference corpus. The software creates this list from a calculation of a value called keyness. This value can be obtained through several different types of calculations, such as the frequently used Log-likelihood and Chi-square (Biber and Reppen,

2015), or in certain cases via Simple Maths, which Kilgarriff (2009) describes as a simple division of occurrences of a word in the focus and reference corpora. Regardless of the calculation used, the resulting keyness value then determines whether they keyword is positive or negative. Positive keywords are those, which occur more frequently in the focus corpus, while negative keywords are the opposite, i.e., those that occur less frequently (Scott, 2009). "Complementary to keywords are lockwords […]. Lockwords are words that occur with similar frequencies in two corpora that we compare" (Brezina 2018).

2.3.1. Aboutness and style

When analysing keywords, it is possible to notice two different types based on their predictability and function in the text. These are keywords related either to the "aboutness" of the text or to its "style." The aboutness could be understood as a term referring to the contents of the text used as a subject of research (Biber and Reppen, 2015). Aboutness, as the name suggests, relates to lexical expressions that describe what the text is about. They include keywords connected to the main concept of the texts used in the corpus, the core topics, or even the author's attitudes (Gabrielatos, 2018). As an example of aboutness, Scott and Tribble (2006) mention a corpus containing Shakespeare's play Romeo and Juliette. If such a corpus would be compared to any corpora consisting of contemporary and everyday English, the keywords obtained as a result of such comparison might contain terms such as *rose*, *love*, *banished*, *death*, or *poison*. Since it is known what the play is about, the presence of such keywords related to the contents of the work of art could be predicted. Such keywords refer to the aboutness of the work.

On the other hand, keywords can also point to the style of the text. This can mostly be observed through grammatical expressions or clusters. Keywords related to style do not help with the summary of the general topic of the texts, but rather provide insight into the author's individual writing style, features of the genre, or the type of texts (Bondi and Scott, 2010). Continuing with the same example of Romeo and Juliette, it is also possible that keywords such as *thou*, *art*, or the pronoun *she* would be present as well. These, however, would not be related to the contents of the play, as their meaning does not reveal any information associated with the story. Instead, they provide information about the style in which the play was written. They are related to the used language and word choice (Scott and Tribble, 2006).

2.3.2. Importance of a reference corpus

Whether or not keywords can be considered useful to the research also depends highly on the choice of a reference corpus. If chosen poorly, it is possible that the findings in the list of keywords may not be as insightful as they would be in a case when a much more suitable corpus would be chosen instead. The attributes that make a reference corpus a "good choice" are, however, largely debatable. The reference corpus should contain similar language as the one used in the primary corpus, and the size of the two corpora should be at least similar (Scott and Tribble, 2006). However, such claims could be disproved by Xiao's and McEnery's (2005) research, in which they compare the effect that reference corpora have on the keyword list by comparing the results of the BNC corpus (100 million words) and Freiburg-LOB corpus (one million words). Their results suggest that there are minimal changes in the two resulting keyword lists; therefore, the size of the reference corpus most likely does not have a significant effect on the keywords found. On the other hand, what Scott (2009) suggests as a more crucial factor when selecting a reference corpus, is the content itself, which should reflect the type of findings we hope to achieve. For example, should one be interested in finding the keywords of the play Romeo and Juliet, a reference corpus consisting of all the Shakespeare plays could be used to eliminate the possibility of Elizabethan English and general information about Shakespeare's idiolect to be present among the resulting keywords (Scott and Tribble, 2006).

2.3.3. Minimum frequency

The final parameter important to consider when examining the list of keywords is the minimum frequency. This parameter is used to cut off words that are considered unusual due to their infrequent occurrence. This can include, for example, proper nouns (Biber and Reppen, 2015). Therefore, should the minimum frequency be set to a value such as 5, the keywords which have occurred less frequently than that would not appear on the list. (Lindquist, 2009). However, even lexical expressions that the minimum frequency would normally filter out can be used for research. This includes even tokens that only exist once in the whole corpus (Biber and Reppen, 2015). These are called "hapax legomena" and could be defined as "words that only appear once in a work of or genus of literature or a body of work by a particular author" (Collins Online Dictionary). Nevertheless, hapaxes are highly localised and therefore evaluating them in the same way as other, much more generalized phenomena could be problematic since the same conclusions cannot be made with hapaxes as with regularly occurring tokens (Biber and Reppen, 2015).

2.4. Collocations

Collocations are, according to Yarowski (1993: 267), "co-occurrences of two words in a […] relationship". These relationships, however, cannot be simply determined by two words following one another. A crucial criterion to consider when discovering collocations is the distance. It specifies the span around the node token (the one we wish to find collocates of) in which other tokens can be considered collocates. This span is called the "collocation window" and its size differs based on the desired results (Brezina, et al., 2015). Collocates do not have to explicitly follow one another to form a collocation. Suppose a software investigating collocations is set to work with a sufficient collocation window, examples (2) and (3) could lead to the same result – that "collect stamps" is classified by the software as a collocation (Lipka, 1992:165).

(2) "They collect stamps."

(3) "They collect many things, but chiefly stamps."

The second and third criteria to consider are frequency and exclusivity, which are closely connected. Frequency, as the name suggests, indicates how many times particular collocates occur together within a text. However, not all such combinations could be called collocations. Exclusivity limits which combinations of lexical expressions could form a collocation based on how exclusive their relationship is. If one of the considered words would form too many other collocations, it is not exclusive enough, and therefore not considered a viable collocate. For example, *love* collocates frequently with *affair*, and neither of the tokens form a high number of other potential collocations. This means that the expression *love affair* has an exclusive relationship and could be labelled as a collocation (Brezina, et al. 2015). While this might be true, however, it is also important to consider the possibility of directionality, i.e., that while *love* collocated frequently with *affair*, the same might not be true vice-versa, and *affair* might not form such a strong collocation with *love* (Gries, 2013).

When dealing with collocations, one might also run into a rather similar phenomenon – colligations. These also deal with words that exist in close relationships, however, while a collocation is a combination of solely lexical expressions, and therefore connected to the lexis, a colligation is related to grammatical words and syntax (Hoey et al., 2007). To put it simply, it is used to "designate the attraction between a lexical item and a grammatical category" (Lehecka, 2015). Staying with *love* as an example, it is possible to find a relationship with the

preposition *in*, forming a colligation *in love* (Brezina, et al., 2015). The same principle can then also be applied to various multi-word phrases, resulting in colligations such as *to the naked eye* (Lehecka, 2015).

Similarly to keywords discussed in chapter 2.3, collocations and colligations can also be measured using several different association measures, which calculate the strength of the relation between the two tokens. Each measurement works differently and therefore produces a different list of collocations. Generally speaking, the association measures work with a combination of frequency of the co-occurring tokens and their exclusivity, i. e., whether or not the two expressions occur predominantly in each other's company (Brezina, 2018). The aim of this paper is not to delve too deep into the calculations of each of the measurements; however, for at least generic summary of the available options, figure 2 shows a graph displaying how different association measures work with collocation frequency and exclusivity.

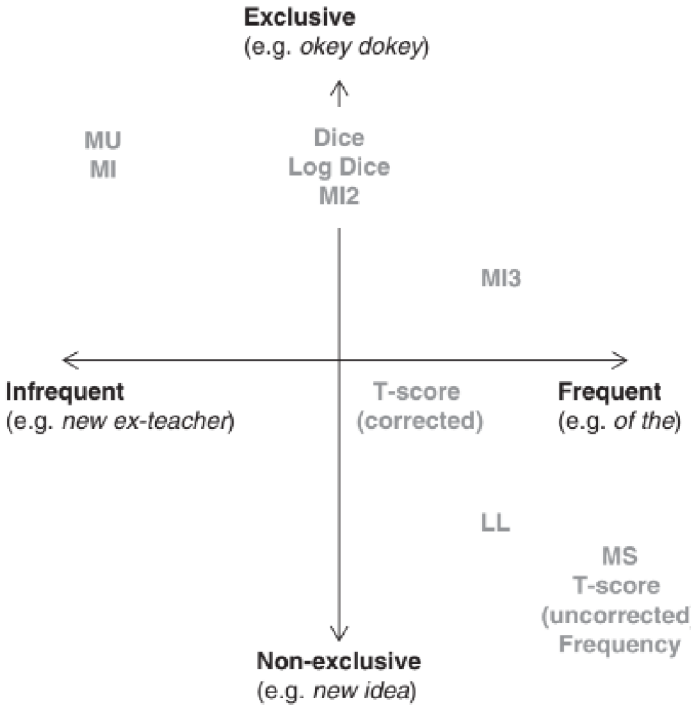

Figure 2: Frequency and exclusivity scale (Brezina, 2018: 74)

For this paper, Log Dice has been chosen as the optimal association measurement due to its attention to the exclusivity of the relations between the collocates. Therefore, all collocations mentioned in the practical part are calculated using Log Dice. The collocation window in all parts of the analysis is set from -3 to 3 unless stated otherwise.

2.5. N-Grams

N-grams, also known as *lexical bundles, recurrent combinations,* or *clusters* (Biber and Reppen, 2015) are strings of several words where the "n" can be substituted by the number of tokens in combination (Lindquist, 2009). In practical research, n-grams are "…the results of the […] procedure which counts continuous multiword sequences and produces frequency lists very much like a word frequency list" (Biber and Reppen, 2015: 45). Clusters are therefore instances of several lexical or grammatical expressions that follow one another, however, unlike in the case of collocations and colligations, only their frequency is taken into account. This means that the resulting table might contain less expected combinations of tokens, which are specific to the given texts (Hyland, 2008). "A word list is therefore similar to a list of n-grams. The simple word frequency list consists of n-grams of length one, but n-grams of length 2, 3, 4, and 5 are usually [also] counted" (Biber and Reppen, 2015: 45). To illustrate how the resulting table of frequencies looks like, the most frequent 5-grams in the BNC were investigated (see fig. 3).

5-gram	*N*
at the end of the	4,531
by the end of the	1,840
i do n't want to	1,761
i do n't know what	1,751
as a result of the	1,597
in the middle of the	1,516
the secretary of state for	1,365
the other side of the	1,234
at the time of the	1,160
you do n't have to	1,050

Figure 3: The most frequent 5-grams in the BNC (Lindquist, 2009: 101)

It is possible to see that the 5-grams include incomplete phrases. This is an example of the frequency of the combination of tokens being the primary focus of the statistic. The computer provides the data for the most common groupings of five co-occurring expressions without taking any other parameters into consideration (Lindquist, 2009). However, should there be the need, the desired results can be specified to, for example, count 2-grams consisting of *good* + common noun. The results can then be seen in figure 4 (ibid.)

Phrase	*N*
good idea	1,851
good news	1,190
good time	876
good thing	822
good deal	783
good job	740
good reason	735
good evening	696
good morning	651
good example	587

Figure 4: Good + common noun in the BNC (Lindquist, 2009: 104)

2.6. Type-token distribution

Type-token distribution (also known as type-token ratio), "expresses the proportion of types (different word forms) relative to the proportion of tokens (running words)" (Brezina, 2018: 57). It is one of the simplest lexical diversity statistics, which is calculated by dividing a number of types in a corpus by a number of tokens in a corpus. Due to the nature of types (described in chapter 2.2), their higher number indicates a more lexically varied text (ibid.). Since the number of tokens in any writing will almost always be higher than a number of types, due to the usual repetition of certain expressions, the result of the calculation will be a number between 0 and 1, as there cannot be fewer tokens than types. The closer the number is to 1, the more types are present in the corpus and the more varied the vocabulary is. The number can then be multiplied by 100 to state the result in a percentage (Biber and Reppen, 2015). The downside of such measurement is, however, the size of the corpus. The longer the measured text, the higher the probability of repeating certain tokens, in which case the token count increases while the type count does not. In such situations, the type-token distribution naturally decreases (Brezina, 2018: 57).

The following text has been used as a sample to demonstrate how type-token distribution is calculated:

> `But what are thoughts? Well, we all have them. They are variously described as ideas, notions, concepts, impressions, perceptions, views, beliefs, opinions, values, and so on. At times they are brief, coming and going in an instant. On other occasions they seem to endure and we can mull them over again and again in the act we call thinking. We can put them aside,

fall asleep, and then return to them later. We refer to them as things we can handle. However, this is just a metaphor." (Williamson, 2009: 1)

When counted the cited text contains 87 tokens, but only 62 types. Therefore, the type-token ratio has been calculated according to the following equation:

A conclusion can therefore be formed, that the short text is highly lexically diverse. 71% of it are different types, and so it contains a rather diverse lexis.

2.7. Drawbacks and limitations of corpus analysis

Since the previous chapters cover the numerous advantages of corpus linguistic analysis, it is important to also discuss its several limitations and drawbacks. One such downside stems from the necessary combination of computational technologies and human insight. While computers may provide assistance and reduce the subjectivity of the analysis, they never truly eliminate it. The raw data collected from any software used for the study of a corpus will inevitably have to be analysed by a researcher to form conclusions (Baker and Egbert, 2016). The final step of an investigation of any topic will always rely on a human element, and therefore, the possibility of an error and faulty assumption is still rather prominent. Despite the major assistance of computational technologies, there is still an ever-present necessity to correctly mediate and interpret the received data (Tognini-Bonelli, 2001). The researcher should, therefore, always pay attention to the information they receive and should never approach the process of forming conclusions with mechanical certainty (Culpeper, 2009).

3. Practical part

3.1. The corpora used

3.1.1. Focus corpus

This section will break down the first steps regarding the preparation of the practical part of this paper. For the purpose of this research, two corpora have been created – the focus corpus, consisting of Terry Pratchett's novel Colour of Magic, and the reference corpus, consisting of seven other fantasy novels released in the same decade – 1980s.

When creating the primary corpus, the first obstacle was the question of dividing the raw text of the book into several text files. This was necessary to be able to view the distribution of the frequent and key words, as well as to be able to investigate any of the texts separately (via filtering), should there be a need. The book contains six chapters. However, two of them are prologues. The first prologue offers insight into the background of the story. It introduces the reader to the setting of Discworld and describes its rather absurd nature (a flat world situated on the back of four elephants, floating through space on the shell of a giant turtle). Due to the potential importance of such introduction, this prologue, while short, has been left intact as a separate text file in the corpus. The other prologue, however, does not have the same function. Instead, it introduces the Gods, which play a rather significant role in the following chapter and then further on towards the end of the story, as they are the ones manipulating the protagonist and other characters. While their introduction could also be labelled as important, it has not been deemed as integral as the general description of the world. The prologue featuring them is more closely tied to the following chapter. Therefore, this part of the book has been merged with the chapter it precedes. In the end, the entire corpus contains 5 texts with over 60 000 words (see table 1)

Table 1: The size of the primary corpus

Primary Corpus	
Tokens	79,888
Words	65,595
Sentences	5, 564
Documents	5

The downside of such division, however, is the uneven distribution of words, since the lengths of the texts vary. An idea of splitting the book into chunks of mostly even token count has been considered, but such a method would almost certainly create difficulties related to the

orientation in the story during the analysis. Therefore, the original division into 5 text files was deemed optimal.

Table 2: Token distribution among the texts of primary corpus

File name	Token count
ColourOfMagic.txt	23,228
Close to the Edge.txt	22,252
The Lure of the Wyrm.txt	19,175
TheSendingOfEight.txt	14,275
Prologue.txt	958

3.1.2. Reference corpus

The reference corpus was created in order to offer a comparison of Pratchett's writing style with that of other writers of the same genre (fantasy). The chosen books were published in the same decade as Colour of Magic, in order to avoid potential difference in language used, which may have been caused simply by the different time of publishing. The candidates for the reference corpus were picked randomly from several online lists, however, a crucial criterion was also the reader rating provided. Since the aim was to compare Colour of Magic to other popular literature of the time, only books of high reader rating were picked for the reference corpus from the websites Goodreads, Amazon, and Google books. The reader rating systems differed from website to website, but the common approach was to not pick anything below 4/5 stars, 80% or 8/10 (with occasional rounding-up, should the value be within a 0.2/10 margin). The books chosen for the reference corpus can be seen in table 3.

Table 3: Token distribution among the texts of reference corpus

File name	Authors	Token count
Daggerspell.txt	Katharine Kerr	189,073
The Drawing of the Three.txt	Stephen King	153,529
Waylander.txt	David Gemmell	120,210
ShadowsLinger.txt	Glen Cook	115,579
The Lives Of Christopher Chant.txt	Diana Wynne Jones	104,868
Howls Moving Castle.txt	Diana Wynne Jones	100,161
Seaward.txt	Susan Cooper	62,509

The information about the reference corpus regarding its size, books included, and their individual word count is provided in table 3 and table 4.

Table 4: The size of the reference corpus

Reference corpus	
Tokens	845,929
Words	692,933
Sentences	61,345
Documents	7

3.2. Software used

Data used in this paper were obtained via Sketch Engine, available online through the institutional access provided by the Technical University of Liberec.

Sketch Engine has been chosen primarily due to its more advanced and user-friendly interface, which offers a more intuitive and quicker access to all the necessary features. Its ability to generate helpful graphical representation of the data has made the research easier and provided figures used later in the practical part of this work.

4. Analysis

4.1. Word list of Colour of Magic

Obtaining a list of the most frequent words has been chosen as the first step of the analysis. The primary reason for this being the nature of the data provided, which are only related to the book (Colour of Magic) and its contents. It has been decided that first the work of art should be investigated on its own before a comparison to other literature would be made.

When obtaining a word list, however, several types of data had to be filtered out, as they most likely would not bring any insight into the analysis. When the word list was first generated without any such specifications regarding the filtration of unwanted results, the most frequent tokens included punctuation marks, such as a dot, a comma, or the quotation mark (see table 5).

These are so called non-words, which are "tokens that do not start with a letter of the alphabet. Examples of non-words are numbers, punctuation but also tokens such as 25-hour, 16-year-old, !mportant, or 3D" (Kilgarriff and Rychlý, 2021).

In the case of this paper, non-words would not be investigated, as in the case of Colour of Magic, the most frequent ones are punctuation, whose presence in the text is unavoidable. In order to remove such results, the option of excluding non-words was used.

Table 5: Word list generated without any filter parameters

Word	Frequency	Word	Frequency
.	4,758	was	1,048
the	4,531	in	986
,	4,003	that	805
“	3,640	said	722
a	1,933	I	722
of	1,924	his	693
and	1,694	you	640
to	1,242	?	629
he	1,164	Rincewind	585
it	1,111	on	523

As a next step, the software was set up to look for lemmas instead of tokens. Because of this, the different forms of the same token would not be taken into account when counting the most frequent lemmas. The reason for choosing this option was mainly to prevent confusion when working with the word list. At the same time, there was always the option to investigate the different forms of a particular lemma later, should the need arise.

However, even after utilizing these three criteria, the wordlist still contained lemmas that were undesirable, as they brought minimal new insight into the research. This includes grammatical words, whose presence in the word list was to be expected, as explained earlier in chapter 2.2. An example of such undesired results were articles, both definite and indefinite, which are generally used often in English language, therefore, their frequent presence does not clearly reveal anything about the contents of the book. Other words that might function in a similar way are certain prepositions, such as *of*, *to*, *in*, *at*, etc., conjunctions, such as *and* and *but*, and verbs *do*, *be*, and *have*. However, this does not mean that grammatical words or punctuation would not have any value. While all of these filtered-out results may still carry useful information, in order to uncover it, more in-depth research would have to be performed, which is not the objective of this paper.

Table 6 shows the final filtered version of the word list that will be used to further analyse the most frequent lemmas. Due to their number, however, only the lemmas whose investigations have been insightful in any way are included in this paper. Some, such as the lemma *it*, are simply used too generically to carry any meaningful information.

Table 6: The finalized word list

Lemma	Frequency	Lemma	Frequency
he	1,164	**would**	212
it	1,111	**by**	211
say	807	**see**	200
that	805	**then**	197
I	723	**its**	195
his	693	**man**	191
you	640	**down**	173
Rincewind	634	**them**	172
not	574	**around**	169
as	431	**so**	168
with	400	**think**	168
Twoflower	349	**like**	165
him	345	**will**	160
there	333	**could**	159
from	280	**go**	156
look	279	**no**	154
for	275	**wizard**	151
they	265	**know**	150
one	259	**dragon**	143
this	254	**which**	143
into	249	**Hrun**	141
out	241	**back**	140
what	240	**about**	136
all	225	**now**	136

Since the list contained a large number of different parts to speech, the results of the finalized word list were first filtered by their POS tags. This was done to make the word list more comprehensible and to help form first theories and questions that would subsequently be investigated and answered.

Table 7: Sorted word list

Noun	Frequency
Rincewind	634
Twoflower	349
man	191
wizard	151
dragon	143
Hrun	141
Pronoun	**Frequency**
He	1,164
it	1,111
I	723
his	693
you	640
him	345
they	265
what	240
its	195
them	172
Verbs	**Frequency**
say	807
would	212
see	200
think	168
will	160
could	159
go	156
know	150

Adverbs	Frequency
not	574
then	197
so	168
now	136
Determiner	**Frequency**
this	254
all	225
which	143
Preposition	**Frequency**
as	431
with	400
from	280
for	275
into	249
by	211
Miscellaneous	**Frequency**
that	805
there	333
look	279
one	259
out	241
down	173
around	169
like	165
no	154
back	140
about	136

The sorted table helps uncover potential relationships between the individual lemmas. In the case of nouns, three of the most frequently used ones are proper nouns, which would point to them belonging to the protagonists. The remaining common nouns then contain *man* and *wizard*. Simply based on this observation, it is possible to form a first question. Do the common nouns in word list describe the characters, to which the proper nouns refer? The answer can be obtained easily by merely reading the book and familiarising oneself with the characters. All three of them are men, and one of them (Rincewind) is a wizard. This would confirm the

question if it was not for the presence of the lemma *dragon*. Neither of the three characters are dragons nor do they travel with a dragon.

In the pronouns section of the table 7, it is possible to notice several instances of masculine types (he, his, him), but no feminine equivalents. Does that mean that the book contains only a low number of female characters? Besides this, *I* and *you* are also present rather frequently, yet the book is not told in first perspective. Does their presence therefore indicate a regular use of direct speech?

A slightly similar phenomenon can be observed in the case of verbs. While most of the listed lemmas in this category occur roughly anywhere between 150 and 200 times, the verb *say* has a raw frequency of over 800. Should direct speech truly be used often in the book, then it might be possible for *say* to immediately follow it, describing which character said what and how.

The miscellaneous group is then used for lemmas that do not fall decisively into a single category, and will most likely have to be examined further to obtain more information. For example, the lemma *look* might be considered a verb describing the act of observation (to look at something), but could also exist as a noun describing an appearance or a facial expression. Which of these is the most common use for *look* in Colour of Magic?

The following chapters will describe how the various lemmas of the word list were investigated and will attempt to answer the questions asked in the previous paragraphs.

4.1.1. Lemmas *he, his, him*

From the word list in table 7, it is clear that the pronoun *he* is the most frequently used lemma (not counting the ones filtered out). In chapter 4.1, it has been mentioned that there are several different masculine pronouns in the word list but no feminine equivalents. This observation might point to a lack of female characters in Colour of Magic.

In order to confirm or disprove this theory, the raw frequencies of *she*, *her* and *hers* were obtained and compared to the masculine counterparts. Upon such investigation, it is safe to confirm that the feminine pronouns are indeed barely used (She = 94 tokens; her = 80 tokens; hers = 1 token). See examples (4) - (6) for context.

(4) *She was the Goddess Who Must Not Be Named; those who sought her never found her, yet she was known to come to the aid of those in greatest need.*

(5) *Several hundred yards away, Liessa was in a strange humour as she strode down the worn steps that led into the hollow heart of the Wyrmberg followed by half a dozen Riders.*

(6) *Hrun turned to Liessa. She shrugged. "Don't I even get a sword?" he pleaded. "A knife, even?" "No," she said.*

To further examine whether or not the story lacks in female characters, a distribution of the word *she* among the texts of the focus corpus has been observed (see table 8). From the results, it is apparent that it is unevenly distributed, and used mostly in the chapter The Lure of the Wyrm. In the case of this chapter, it might be speculated that lemma *she* would often refer to Liessa Wyrmbidder, a female ruler or Wyrmberg, a location where this particular part of the story takes place.

*Table 8: The frequency of the lemma `she_*

File name	Raw Frequency
The Lure of the Wyrm.txt	46
Close to the Edge.txt	24
TheSendingOfEight.txt	22
ColourOfMagic.txt	2

After investigating the lemma *she* in the context (using KWIC) of the Lure of the Wyrm, it is clear that most of the pronouns indeed refer to the ruler of Wyrmberg. However, this further proves that there is no other major female character in this chapter. Further research has been carried out in to see if the pronoun refers only to a single character in the other texts as well. This is only partially truthful, as the lemma *she* from the chapters Close to the Edge and The Sending of Eight nearly always refers to the same female character – a goddess who cannot be named, which is also a reason why the word *she* is used so often to describe her, as she lacks any name that could be used instead. In the chapter Colour of Magic, there is no major female character, therefore, the pronoun is barely used.

From these findings, it is safe to assume that the book contains a much larger variety of male characters, as there are only a few mentions of any other women, except for Liessa and the goddess.

4.1.2. Lemmas *say*, *I* and *you*

The reason for the lemma's presence in the word list became apparent from the investigation of its context. It is primarily used in the past tense *said*, following a direct speech, as seen in

example (7). To investigate this phenomenon further, a simple calculation has been carried out to determine the approximate percentage of cases in which the lemma *say* is used only in the past tense and after direct speech. For this calculation, Sketch Engine's shuffle feature[1] has been used to create a mix of 100 semi-random examples of the lemma used in context. From these results, it has been calculated that 82% of all instances of the lemma *say* have been used in past tense after direct speech – examples (7) – (10). The remaining 18 per cent include the types *saying* and *says/say*, along with the past tense *said* not used after direct speech. This would indicate that the lemma is indeed used mostly for monologues and dialogues.

(7) `I saved us from [illegible] said Twoflower, remember,

(8) *"Pleased to meet you," said Rincewind.*

(9) *He's doing all right on his own," said the innkeeper, but took a few steps backward.*

(10) *Hrun was said to be roving somewhere Turnwise.*

To further explore how frequently the author uses direct speech, the aforementioned pronouns *I* and *you* were also investigated in context, just as lemma *say*. The shuffle feature provided 100 pseudo-random KWIC samples, and it has been observed that both *I* and *you* are connected exclusively with direct speech. The calculations have then been verified, due to their exceptionless nature, by generating another 100 different samples with a second use of shuffle. Nevertheless, the results have the remained the same. Therefore, it is safe to state that these lemmas are only used in direct speech.

Based on the findings described in the previous two paragraphs, it may be possible to assume that Pratchett's writing style relies heavily on dialogues. This can be seen via the raw frequencies of *say*, *I*, and *you* (see table 9, 10 and 11), which are connected with direct speech in a vast majority of cases.

(11) *"Look," said Rincewind, "this isn't getting us anywhere."*

(12) *"Sometimes I think a man could wander across the disc all his life and not see everything there is to see," said Twoflower.*

---

[1] Shuffle feature reorganizes the examples in KWIC in a semi-random order. Applying the tool on the same concordance will always produce the same new order of lines. This is intentional so that other researchers may reproduce the results. Applying the shuffle tool more times will also always produce the same results.

(13) *"You were supposed to be on watch," snapped Rincewind. "I saved us from the slavers, remember," said Twoflower.*

As the next step, the distribution of the lemmas among texts has been investigated to see if they have perhaps been used more in a specific chapter, which could point out a particular abundance of dialogue. However, the lemma *say* appears to be relatively evenly distributed among the texts, as seen in table 9 with its relative frequencies. The slight differences in the numbers might be attributed to the different natures of individual chapters. Since the main characters briefly split up in Lure of the Wyrm and The Sending of Eight, there is a slightly lower amount of dialogue in these parts when compared to those where the characters travel together. The minimal occurrence of the lemma *say* in prologue then confirms these statements, as there is no dialogue at all in this chapter. It consists of a monologue of the narrator directed at the reader, therefore there is no direct speech. Lemmas *I* and *you* then have similar distributions (see tables 10 and 11).

Table 9: The frequency of the lemma "say"

File name	Raw frequency
ColourOfMagic.txt	268
Close to the Edge.txt	237
The Lure of the Wyrm.txt	180
TheSendingOfEight.txt	121
Prologue.txt	1

Table 10: The frequency of the lemma "I"

File name	Raw frequency
ColourOfMagic.txt	230
Close to the Edge.txt	223
The Lure of the Wyrm.txt	176
TheSendingOfEight.txt	94

Table 11: The frequency of the lemma "you"

File name	Raw frequency
ColourOfMagic.txt	232
Close to the Edge.txt	177
The Lure of the Wyrm.txt	154
TheSendingOfEight.txt	77

The theory that the lemma *say* is used mainly after direct speech in the book is further proven by investigating the collocations. From the resulting visualization (see fig. 5), it appears evident

that the lemma in question collocates most frequently with the two main protagonists, Rincewind and Twoflower. This makes sense since these characters appear often in the book and would therefore take part in most dialogues. However, there are also other names and nouns that appear in the figure, which point to other characters in the story, such as a troll, Hrun, patrician, Weasel, imp, etc., also having their dialogues.

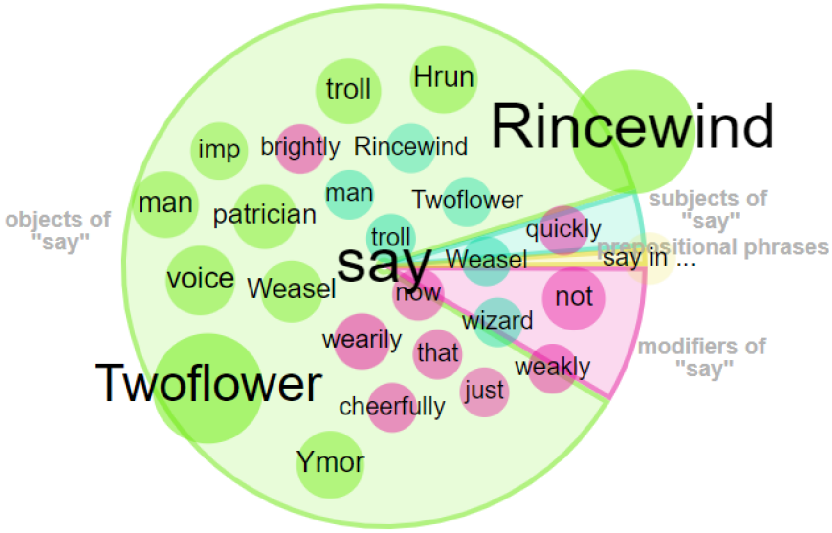

Figure 5: Collocations with lemma "say"

As seen in table 7, different types of the lemma *he*, such as *his* or *him* are also present among the most frequent words. Their feminine equivalents have already been investigated in chapter 4.1.1. However, similar types of lemmas *I* and *you* are not on the word list. It has been stated that the book features frequent dialogues and direct speech, so how come the lemmas *me*, *my*, *mine*, *your*, and *yours* are not used as often? Table 12 shows the raw frequency of each of the aforementioned expressions. It can be seen that while they do exist in the book, their number is significantly lower than that of the corresponding lemmas.

Table 12: Raw frequencies of types of lemmas "I", "you", and "he"

Type	Raw frequency
me	117
my	90
mine	3
your	74
yours	7
his	613
him	345

When a similar investigation has been performed with the types *his* and *him*, it has been observed that they occurred outside of direct speech as well, appearing also in descriptions. For comparison, see examples (14) – (17).

(14) *"If you will follow me to my house I will find you food and a change of clothing," said the troll solemnly.*

(15) *"And now I really think I should introduce myself. Why has your friend gone that strange colour?" "Culture shock, I imagine," said Twoflower.*

(16) *The Arch-astronomer of Krull motioned lightly with his hand and his bearers set the throne down in the shadow of the hull.*

(17) *Twoflower snored on. Rincewind jabbed him viciously in the ribs. "I said, are you awake?"*

Generally speaking, the frequency of *his* and *him* can be explained by the fact that they are simply used more often due to their possibility to be present outside of direct speech. The same, however, cannot be said about the other lemmas featured in table 12. Those only occur in direct speech, and therefore less often than their masculine counterparts.

4.1.3. Lemma *Rincewind*

Due to the fact that this character is the main protagonist of the story, it makes sense for his name to be among the most frequently used lexical expressions. Since such a character should most likely have a memorable and unique personality, a question arises: Is it possible to observe and describe a specific lexis the author used when describing this character and his actions?

The distribution of the lemma among the texts is not entirely even, which might seem unusual for a protagonist of the book, however, upon closer investigation, the data have a logical explanation behind them. As seen from table 15, lemma *Rincewind* is not used at all in the prologue. This is only logical, since the prologue introduces the world itself and not any particular character. The chapter Lure of the Wyrm features a significantly lower number of instances of the name, which can be attributed to the fact that a part of the chapter focuses on Twoflower, the secondary protagonist. Therefore, the focus briefly shifts away from Rincewind.

Table 13: The frequency of the lemma "Rincewind" among the texts ˉ relative frequency per million

File name	Raw frequency	Relative frequency
Close to the Edge.txt	193	8,673.38
ColourOfMagic.txt	193	8,308.94
TheSendingOfEight.txt	138	9,667.25
The Lure of the Wyrm.txt	110	5,736.64

The collocates with the lemma *Rincewind* offer an interesting insight into the way the author describes the character's personality without explicitly stating what Rincewind is like. The strongest collocate is the lemma *say*, which has been investigated earlier in the paper (See chapter 4.1.2.), therefore this serves as a confirmation that the author truly uses the verb *to say* with direct speech. This is apparent from figure 5, where we can see how strongly the lemma *say* co-occurred with Rincewind. However, since it might be considered natural for characters to say something in a specific manner, usually described by the use of adverbs, the co-occurrence of *said Rincewind* has been looked into further to see if the author perhaps had a tendency to depict in what way would the character say something. From the list of collocations, it is possible to notice the presence of adverbs *hurriedly*, *urgently*, *gloomily*, *miserably* or *irritably* (see examples 18 – 22), which are usually used to depict the protagonist's reactions to various situations. In this case, the adverbs present are mostly negative in meaning, which could be considered fitting for the character since the protagonist is often depicted as grumpy, cowardly, pessimistic, and easily annoyed.

(18) *"They don't look very roomy to me," said Rincewind hurriedly, and grabbed the tourist by the arm, "so if you'd just come on, no sense in staying here-"*

(19) *The weight swung away, pulling a pin from an intricate little mechanism. A chain began to move. There was a clonk... "What was that?" said Rincewind urgently.*

(20) *ŭthe eight colours of the rainbow reflected in the many telescope lenses of the city's multitude of astronomers. "It's absolutely awful," said Rincewind gloomily.*

(21) *You will be a guide, Rincewind, to this looker, this Twoflower. You will see that he returns home with a good report of our little homeland. What do you say to that?" "Er. Thank you, lord," said Rincewind miserably.*

(22) *The imp tapped the side of the box meaningfully. "We'll see who sinks first." The luggage yawned, and moved forward a fraction of an inch. "Oh, all right," said Rincewind irritably.*

Other verbs that collocate with *Rincewind*, are then also mostly corresponding to the character's personality. Lemmas such as *snap*, *mutter*, *snarl*, *moan* or *scream* all fit his description, which was mentioned above. In conclusion, these findings show that the author is trying to use these diverse verbs and adverbs to illustrate a particular behaviour that is typical for the character. A slightly different phenomenon can then be observed with collocations that include Rincewind as a subject. They mostly include verbs that describe the character's senses and perception of the world, such as *look*, *see*, and *stare* for visual observation, *think* and *know* for mental processes, and *feel* for emotional states and frames of mind (see examples 23 - 25).

(23) *Rincewind looked at the shimmering fists that rested lightly in the troll's lap. He suspected they could strike with all the force of a tsunami.*

(24) *Rincewind thought that a meeting with most of the Drum's clientele would mean that Twoflower never went home again, unless he lived downriver and happened to float past.*

(25) *Rincewind felt his will draining away like water from a sieve.*

Other insightful collocations are those used with the possessive Rincewind's. The strongest ones always describe the character's body part and actions/feelings associated with them (examples 26 – 28).

(26) *Light dawned inside Rincewind's head.*

(27) *Agony shot up Rincewind's arm.*

(28) *Rincewind's elbow nudged something. It was the tree trunk.*

In conclusion, it is possible to observe that Pratchett's idiolect when describing this character tends to be very specific. When the character appears as the object of the sentence, the author uses verbs associated with their personality and connected with actions noticeable on the outside. However, when the character's name is used as a subject of the sentence, what usually follows is a verb describing their senses and feelings, which are purely internal. Additionally, the author frequently uses specific body parts of the character when writing about their actions.

For example, instead of saying that *Rincewind nudged something*, he specifies the situation further by writing *Rincewind's elbow nudged something*. Can the same be observed with other characters as well?

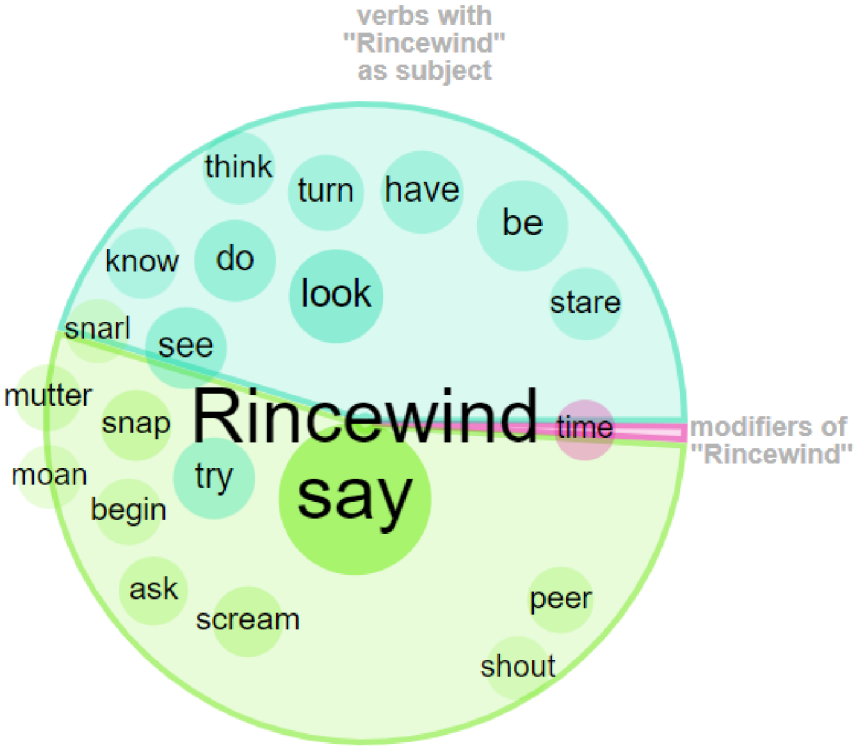

Figure 6: Collocations with lemma "Rincewind"

4.1.4. Lemma *Twoflower*

The lemma Twoflower should not significantly differ in its analysis from the lemma Rincewind due to the character playing the role of a secondary protagonist. To either prove or disprove such expectation, the collected data were always compared to Rincewind in order to spot potential differences.

The first measured parameter was the distribution of the lemma among the texts, which offered expected results. As seen in table 16, the relative frequencies are always lower than in the case of Rincewind (table 15), which would correspond to the fact that Twoflower is the secondary protagonist. The biggest difference between the two then lies within the chapter Colour of Magic. Here, however, Twoflower's lower relative frequency can be justified by his introduction, which only happens about halfway through the chapter, therefore, it only makes sense that the name of Rincewind, who was present in the story since the beginning, occurs more often in this part of the book.

Table 14: The frequency of the lemma "Twoflower" among the texts ˜ relative frequency per million

File name	Raw frequency	Relative frequency
Close to the Edge.txt	116	5,213.01
The Lure of the Wyrm.txt	89	4,641.46
TheSendingOfEight.txt	73	5,113.84
ColourOfMagic.txt	71	3,056.66

The next criterion to consider were the collocations. Here, the main objective was to answer the question from chapter 4.1.3; whether or not one can use collocations to observe a similar usage of language as in the case of Rincewind. The strongest collocate of Twoflower is *say* as well. Therefore, based on the findings described in the previous chapter about the collocation *said Rincewind*, a question could be asked: does *said Twoflower* also co-occur with adverbs that would correspond to his personality? Are they more positive or negative? Upon investigating the adverbs that follow the co-occurrence *said Twoflower*, expressions such as *comfortingly*, *cheerfully*, or *excitedly* were found (see examples 29 – 31). Their meaning is positive, unlike in the case of Rincewind. However, such adverbs are expectable in the case of Twoflower, as he is depicted as more optimistic, carefree, and excited about the world around him. Therefore, it could be said that in Colour of Magic, Pratchett frequently used the lemma *say* followed by adverbs that would fit the personality of the speaking character.

(29) *"Look, I'm sorry I steered us into the reef, but this boat doesn't seem to want to sink and we're bound to strike land sooner or later," said Twoflower comfortingly.*

(30) *"Well, off again then," said Twoflower cheerfully. He turned and waved at the troll, now no more than a speck on the edge of the world.*

(31) *"Are you a goddess then?" said Twoflower excitedly. "I've always wanted to meet one."*

Subsequently, the next strongest collocations with Twoflower as an object of the sentence include verbs related to actions, such as *to ask*, *to explain*, or *to wail*. In this regard, the results are similar to those of Rincewind. The verbs describe what the character's typical behaviour is like. As a tourist, Twoflower often asks about various unknown subjects or explains how different they are from his homeland. The *to wail* is then used to express the character's frustration with facing the unknown.

(32) *"When it gets really dark, do you think we'll be able to look down and see Great A'tuin the World Turtle?" asked Twoflower, staring at the rolling clouds.*

(33) *"I don't think you understand," explained Twoflower. "I am a citizen of the Golden Empire. I'm sure Krull would not wish to incur the displeasure of the Emperor."*

(34) *"But you haven't actually told us anything!" wailed Twoflower. "It's not really worth it, is it? What with you being sacrificed in the morning," said Garhartra.*

Next, the collocates with Twoflower as a subject were also observed and here, the results slightly differ from those obtained in the investigation of Rincewind. While it is true that verbs describing internal processes and feelings, such as *to think* are still present, other predicates, related to external actions are also included. Examples of such verbs could be *to sit*, *to follow*, *to watch*, and *to add*, the latter being tied to dialogues.

(35) *At one point Twoflower thought he heard shouts, a long way behind them but they soon faded away.*

(36) *Twoflower sat down on an ornate mother-of-pearl chair with a glass of oily wine in one hand and a crystallised squid in the other.*

(37) *"All that gold and silk and stuff," Twoflower added. "I've never seen anything like it!"*

Finally, the possessive Twoflower's reflect the findings of Rincewind, s the lexical expressions are related to body parts (Twoflower's face, Twoflower's arm), however, there are several exceptions, which were not present in the case of Rincewind. Some of the strongest collocations with the aforementioned possessive are *horse*, *voice*, and *laundry*. While *voice* could theoretically be counted among the lemmas expressing body parts, since it is a sound created by a body, *horse* and *laundry* are of a completely different category, referring to a mode of transportation and clothing. Interestingly, nothing similar has been observed with Rincewind, despite the fact that he rode a horse as well and his apparel was also described numerous times.

(38) *The creature grunted, and with one swipe of a claw gouged a wound in the flank of Twoflower's horse, which screamed and bolted into the trees at the roadside.*

(39) *Savagely he wrenched the lid up. There was nothing inside but Twoflower's laundry. It was perfectly dry.*

In conclusion, it is impossible to say that the author would always stick to the same habits regarding collocations as described in chapter 4.1.3; however, the comparison does not deny it. The findings are not identical but they still point to numerous similarities that should not be ignored. Therefore, it might be safe to state that Pratchett's idiolect in Colour of Magic does indeed contain a tendency to write certain types of lexical expressions in relation to characters, however, such phenomenon is not always present and therefore does not directly define the author's used lexis.

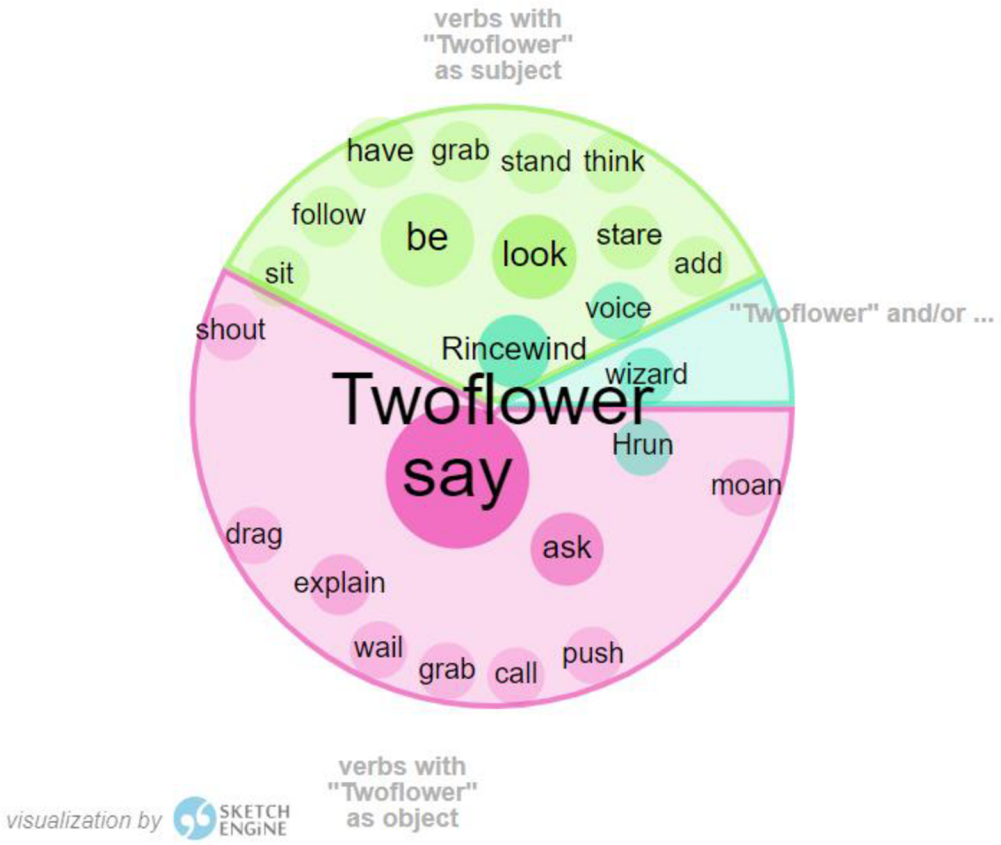

Figure 7: Collocations with lemma "Twoflower"

4.1.5. Lemma *dragon*

As mentioned in chapter 4.1, the lemma *dragon* is among the most frequently used nouns, despite the fact that none of the main characters are dragons nor do they travel with such a creature. Therefore, why is the lemma *dragon* used so frequently?

Firstly, the aim was to verify that the lemma is used evenly throughout all the chapters, which, as seen in table 17, is not the case. *Dragon* is present almost exclusively in the part Lure of the Wyrm. What might be seen as unusual is that the raw frequency of *dragon* in Lure of the Wyrm is higher than that of both protagonists. While it is true that the mythical creatures are present throughout the whole chapter, it might be considered unusual for them to occur more often than the name of those the entire book is about. Therefore, why is it that the lemma *dragon* is used more often in this chapter than the protagonists' names?

Table 15: The distribution of the lemma "dragon" ˜ relative frequency per million

File name	Raw Frequency	Relative frequency
The Lure of the Wyrm.txt	139	7,249.02
Close to the Edge.txt	2	89.88
TheSendingOfEight.txt	2	140.11

The answer was discovered while investigating the text in a larger context and re-reading the chapter. Certain common nouns are often used to refer to the main characters, such as *wizard* in the case of Rincewind or *tourist* in the case of Twoflower. At the same time, the names can also be replaced by *man* or *men*. When it comes to *dragon*; however, there are not many available synonyms. Usually, the pronoun *it* has been used as a proform of *dragon*, and sometimes, it is possible to observe the noun *creature* being used instead. Nevertheless, presumably since the latter can also refer to countless other living beings, the author seems to avoid it and prefers to instead use the lemma *dragon* to describe the beings specifically. Thesaurus has also been consulted in an attempt to find any other possible synonyms. Still, these were limited to terms such as *basilisk*, *hydra*, or *wyvern*, neither of which are present in Colour of Magic.

In conclusion, it might be safe to say that the lemma *dragon* is used so frequently because of the lack of better lexical expressions that could be used to accurately describe the same creature. In the case of the protagonists, there are more options.

Just as in the previous chapters, the most frequent collocations of the term *dragon* were also investigated. In this case, the findings were less unexpected than with the protagonists. Verbs co-occurring with *dragon* as an object usually refer to actions performed by other characters. This can be seen in example (40). Verbs with *dragon* as a subject then refer to the physical actions of the creatures, such as *to fly*, *to roar*, or *to rise*. Possessive d r a g o n ˇ s collocates exclusively with different body parts, including d r a g o n ˇ s

chest. These are then sometimes accompanied by further specifications of the exact area of the dragon's body, as depicted in example (42).

(40) *"All my life I've wanted to see dragons!" "From the inside?" shouted Rincewind.*

(41) *As the dragon rose higher above the patch of woodland, where the three of them had slept a damp and uneasy sleep, the sun rose over the edge of the disc.*

(42) *Again, the ball of flame rolled out but this time, as the dragon's neck muscles contracted, its colour faded from orange to yellow, from yellow to white, and finally to the faintest of blues.*

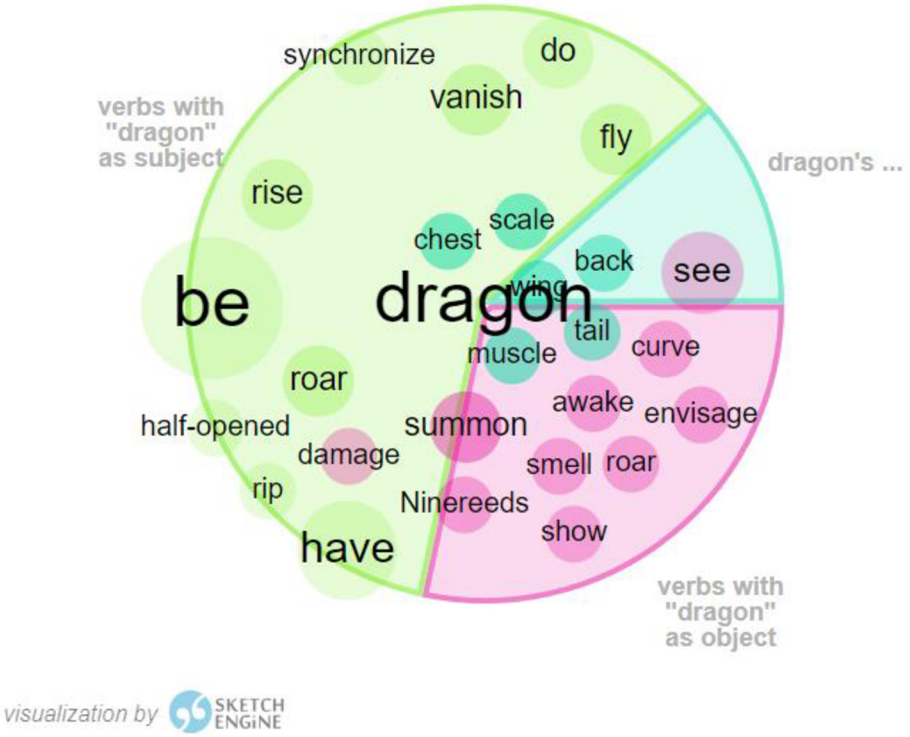

Figure 8: Collocations with lemma "dragon"

For comparison, the same lemma has been investigated in the 80s fantasy literature corpus as well. The objective was to discover whether or not dragons were used frequently in other popular fantasy book of the decade. The results have pointed out that the raw frequency of the lemma in the reference corpus is 71, around half of what can be found in Colour of Magic. The frequency of the expression among the individual books can be seen in table 18. There, it is shown that the lemma exists in almost all the reference literature, albeit often with very low raw frequency. Therefore, it could be said that Colour of Magic focuses on these mythical creatures a lot more than the reference literature.

Table 16: Raw frequency of lemma "dragon" in the reference books

File name	Raw frequency
The Lives Of Christopher Chant.txt	36
Seaward.txt	18
Daggerspell.txt	12
ShadowsLinger.txt	2
Waylander.txt	2
The Drawing of the Three.txt	1

4.1.6. Lemma *as*

This particular lemma is evenly dispersed among the texts. It is often used very generically in colligations such as *as many as*, *as far as*, *as heavy as*, etc. It also frequently collocates with the lemma *Rincewind*, which is expectable, given that the character is the protagonist of the book. What could be seen as a rather unusual finding; however, is the presence of the lemma *dragon* among the strongest collocates. Not only that, but according to the LogDice calculation, the collocation formed with *dragon* has the value of 9.58, which is higher than in the case of *Rincewind* (LogDice = 9.47). From table 15 in chapter 4.1.5, it is possible to see that *dragon* is used almost exclusively in the chapter Lure of the Wyrm and its high frequency may be attributed to the lack of a suitable proform or synonym. Therefore, it might be possible to assume that the colligation *as* + *dragon* has such a high LogDice value for the same reason. However, such idea has been partially disproved by investigating the coocurence in context using KWIC. It has been discovered that the author uses this particular combination of lemmas almost exclusively in the following pattern: *as* + *DT/ PPZ* + *dragon* + *VVD* (see figure 9). This, combined with past findings, led to the formation of a theory: Did the author have several phrases which he was accustomed to and therefore used without much variation? In order to answer this question, a similar pattern was searched for as part of the investigations of the other lemmas from the word list. At the same time, an analysis of *as* has been performed in the reference corpus to look for cases of a similar pattern; however, only a very few have been found. In Howl's Moving Castle, there are a few instances of a s t h e which only makes sense due to the setting of the book. Another example could be the phrase *as the days passed*, which is; however, too generic. It does not refer to a particular place, creature, or a character, therefore, it does not provide much evidence for the question. This would therefore point to the fact that in Colour of Magic, Pratchett would use certain established phrases. To further prove this; however, more evidence has to be gathered.

as	the **dragon** veered and sped c
as	the **dragon** rose out of its long
As	his **dragon** swooped away Lio
as	Liartes' **dragon** thundered by,
as	the **dragon** circled slowly, tiltin
As	the **dragon** rose higher above
As	the gold beast materialised in t
as	the **dragon** ripped its way thro
As	the **dragon** banked slightly he

Figure 9: KWIC examples of the collocation: `as ŭ dra

4.1.7. Lemma *with*

In the case of this lemma, a similar structure as with the colligation a s ŭ d r a g o n was observed. Upon investigating the collocates and sorting them in a descending order by their LogDice value, a strong co-occurrence with the lemma *interest* has been identified. This has been once again examined in context and a pattern similar to the one discussed in the previous chapter has been found. All of the instances of this colligation are connected to words with meaning related to sight, such as *to watch*, *to look*, *to peer*, or *to observe*. The pattern then always looks like this: *somebody* + *verb related to sight* + *with interest* (see figure 10). This could be labelled as another structure which the author is accustomed to and uses automatically with little variation, similar to the previously mentioned pattern with *as* + *DT/PPZ* + *dragon*. However, these two findings are not sufficient enough evidence to definitively prove such a statement. In order to do so, much more extensive research would have to be carried out, which would perhaps look into other works by Terry Pratchett and try to find phrases of similar nature.

The colligation was once again investigated in the context of the reference corpus as well. This time, numerous instances of the same pattern have been discovered, however, only seven can be detected in the whole reference corpus (see figure 11). This could serve as evidence further indicating that patterns such as *as* + *DT/PPZ* + *dragon* or *somebody* + *verb related to sight* + *with interest* might be typical for Pratchett, at least in the context of Colour of Magic. At the same time, it is possible to see that Pratchett used the phrase *with interest* without any

premodification of the noun *interest*. On the other hand, the examples from the reference corpus almost always contain a certain premodification of the lemma *interest*. Compare figure 10 and figure 11.

;ymbols, stood aside and watched	with	**interest** as the sextet passed.</s>
, where two figures were watching	with	considerable **interest** .</s><s>The
ınce gawpers were watching them	with	**interest** .</s><s>"Look," he said, "
ind, and then peered at Twoflower	with	**interest** .</s><s>"Everything all rig
new-found admirers watching him	with	**interest** in case he did something I
dbye."</s><s>The crowd watched	with	**interest** .</s><s>It was turning out
ıearby stall watched this madman	with	**interest** .</s><s>I COULD LEND Y
ing on its lid watching the scenery	with	**interest** .</s><s>Bravd looked at t
ned, but the dragonrider observed	with	some **interest** the strange way in v
nen were watching Psepha's flight	with	**interest** .</s><s>Rincewind swallc
goyles.</s><s>Their eyes glowed	with	**interest** .</s><s>"Attention, please
ınd the base of the tree looked up	with	**interest** at their next meal talking t
ceably thinner, were watching him	with	**interest** .</s><s>"The salamander

Figure 10: KWIC examples of the in the Colour of Magic corpus `with interest

Left context	KWIC	Right context
as outside?"</s><s>Michael asked	with great interest	.</s><s>But Howl came cha
Miss Angorian and were watching	with interest	to see what would happen.<
t this, where Calcifer was watching	with some interest	.</s><s>"He isn't!"</s><s>S
debt, stealing food from his mouth	with ridiculous interest	rates.</s><s>On the other h
ck a small crowd that stared at him	with uneasy, avid interest	as he was led away.</s><s>
ed, looking at the men around him	with amused interest	, as if he wasn't going crazy
had read of the Parks incident, but	with little interest	at first.</s><s>That came lit
d her, but the gunslinger observed	with some real interest	that she went on speaking f
>Christopher peered down inside it	with great interest	.</s><s>The wound was a r
do you make it?"</s><s>he asked	with great interest	.</s><s>Before the Goddes
ds.</s><s>Christopher looked at it	with interest	, wondering what it had to d
</s><s>Christopher listened to this	with some interest	, because he had always wo

Figure 11: KWIC examples of the collocation with interest in the fantasy literature corpus

4.1.8. Lemma *by*

The data related to this grammatical word have been difficult to analyse. Its distribution is highly uneven and the lemma occurs with a high relative frequency in the prologue (see table 19), which is the opposite of usual findings. Due to the prologue's small word count, most lemmas investigated do not occur as frequently in the prologue as in other chapters, even when comparing relative frequencies. Therefore, to find such a high relative frequency in the prologue was unusual.

Table 17: The frequencies of the lemma "by" ˜ relative frequency per million

File name	Raw Frequency	Relative Frequency
ColourOfMagic.txt	68	2,927.50
Close to the Edge.txt	48	2,157.11
The Lure of the Wyrm.txt	46	2,398.96
TheSendingOfEight.txt	40	2,802.10
Prologue.txt	9	9,394.57

Upon investigating the lemma in context, the reason for this phenomenon became clearer. There are only nine occurrences of lemma *by* in the prologue, but they all appear to contribute to the description of the world. The patterns observed usually consist of a passive voice in which the

word *by* introduces either a part of the world (as seen in example 43) or a particular group of people (example 44).

(43) "ŭ g a r l a by the long *waterfall at its vast circumference and domed by the baby-blue vault of Heaven*."

(44) `*An alternative, favoured* ***by*** t h o s e o f a r e l i g i o u s p e r s u

In conclusion, the frequent usage of the lemma *by* in the prologue might be attributed to the natural occurrence of the preposition in passive voice, which was the author's preferred voice when describing his world.

; Weight.</s><s>Most of the weight is of course accounted for	by	Berilia, Tubul, Great T'Phon and
d star tanned shoulders the disc of the World rests, garlanded	by	the long waterfall at its vast circ
ıded by the long waterfall at its vast circumference and domed	by	the baby-blue vault of Heaven.<
>The early astrozoologists, hauled back from their long dangle	by	enormous teams of slaves, wer
as popular among academics.</s><s>An alternative, favoured	by	those of a religious persuasion,
vere all the stars in the sky which were, obviously, also carried	by	giant turtles.</s><s>When they
ime to its first midsummer (Small Gods' Eve) which is followed	by	Autumn Prime and, straddling tl
vel at its heart.</s><s>Since the Hub is never closely warmed	by	the weak sun the lands there ar
ccult significance on the disc and must never, ever, be spoken	by	a wizard.</s><s>Precisely why

Figure 12: the lemma "by" displayed in the context of the prologue

4.1.9. Lemma *look*

Since the lemma was categorised among the miscellaneous section of the word list, which has been explained in chapter 4.1, the primary objective was to identify which part of speech the lemma represents most of the time. This was discovered by using KWIC and displaying only one-word class at a time. The results show that only 22 tokens correspond to a noun, while in the remaining 257 cases, the lexical expression is used as a verb. Therefore, it might be safe to place it among the other verbs of the word list, as only about 8% of the tokens are of a different word class.

As with other verbs, collocations of *look* have been examined. The outcome; however, was mostly predictable. The subjects of *look* nearly exclusively relate to characters, with Rincewind and Twoflower forming the strongest collocations. The modifiers of *look* then often fit the particular character or situation. There has been an attempt at investigating the adverbs further

by splitting them into categories by their meaning (positive/negative); however, each of the -ly adverbs seen in figure 13 occurs only once or twice in the whole corpus, and they are often connected to different characters, not only the protagonists. Therefore, with such few examples and such a high variety of characters, it is not possible to form conclusions.

Look is, however, used in the book mostly as a phrasal verb, forming colligations such as *look around*, *look down*, or *look out*, which might also be considered a predictable find.

(45) *Twoflower and Hrun looked around the little hollow where they had made their noonday halt.*

(46) ŭ t h e B r o k e n D r u m *looks down* s t h e p l a n a c r o s s t h e s t r e t h a t i t

(47) *Rincewind looked out into the phosphorescent darkness that surrounded the island, and grinned inanely.*

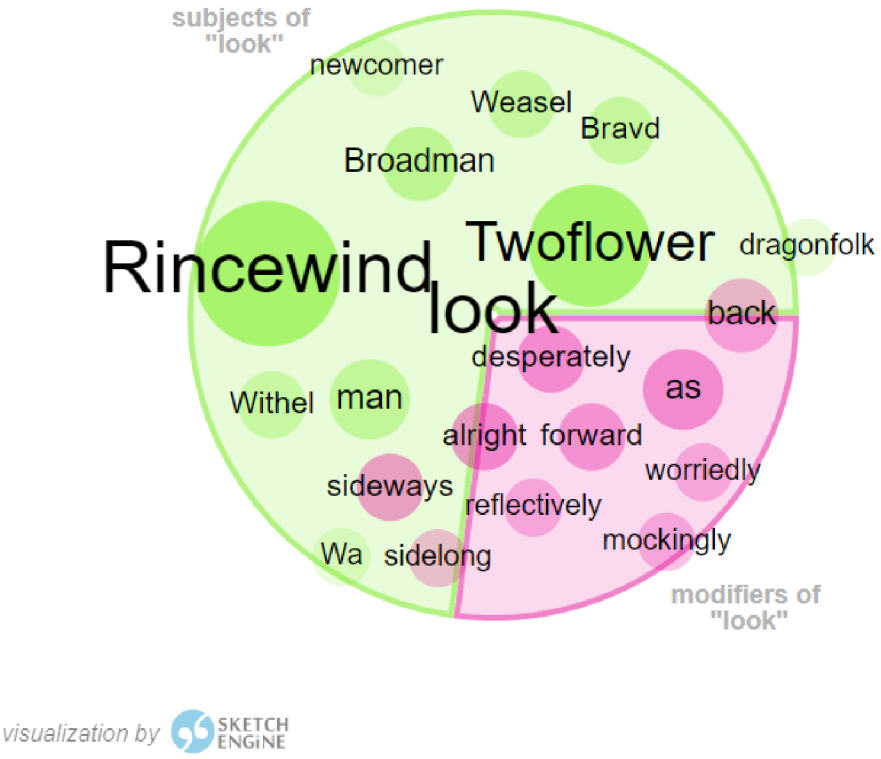

Figure 13: Collocations with lemma "look"

4.2. N-grams

Similarly, as done in chapter 4.1, a list of N-grams has been created and the most frequent N-grams were explored to uncover whether their presence would help prove or disprove some of the aforementioned questions seen in the sub-chapters of chapter 4.1.

The first step was the creation of the list and the decision of how large would the investigated clusters be. Since 2-grams would most likely include generic results, such as *do not*, or various combinations of prepositions + articles, which would not be very insightful, the size 3 was chosen for the first list. Subsequently, a list of 5-grams was also be generated to look for differences between the two.

Table 18: 3-grams of Colour of Magic (lemmas)

Lemma	Raw frequency
There be a	77
I do not	49
Out of the	43
It be a	36
One of the	35
In front of	31
The edge of	26
Look at the	23
Over the edge	22
A number of	22
There be no	22

Table 18 shows that even with 3-grams, one does not necessarily avoid constructions such as *preposition + article*, which were mentioned in the previous paragraph. Such result can be seen in the form of *out of the* or *in front of*, which even upon investigating in context provide a very mixed series of results.

The presence of certain items on the list is then predictable due to the nature of the story. *Over the edge* or *the edge of* are usually referring to the edge of the Discworld, which, as the title of the chapter Close to the Edge implies, is a setting of the finale of the novel. Therefore, a presence of such clusters could be expected. The same could then be said about *look at the*. Since chapter 4.1.9 explained that the verb *look* is mostly used along with characters to describe what they see, the cluster *look at the* would naturally occur rather often.

The two most insightful items of the list are *there be a* and *I do not*, which can be seen in examples (48) and (49). When examined in context, the first item is connected with sentences that describe events and time. Although the reason of its presence was not obvious at first, it appears logical for such a construction to be among the most frequent clusters. Descriptions such as the ones using the tokens *There was* a ŭare to be expected in a book. What appeared as unexpected, on the other hand, was the discovery that *there be a* is rather frequently followed

by lemmas expressing phenomena related to light or sounds. Examples taken from KWIC can be seen in figure 14. The latter of the two clusters, *I do not*, then serves as further evidence of a frequent usage of direct speech as all the instances for this n-gram occur within the inverted commas of a direct speech.

(48) *There was a crackle of octarine flame from his fingers and the air suddenly took on the thick, greasy feel that indicated a powerful magical discharge.*

(49) *"I don't think you understand," explained Twoflower. "I am a citizen of the Golden Empire.*

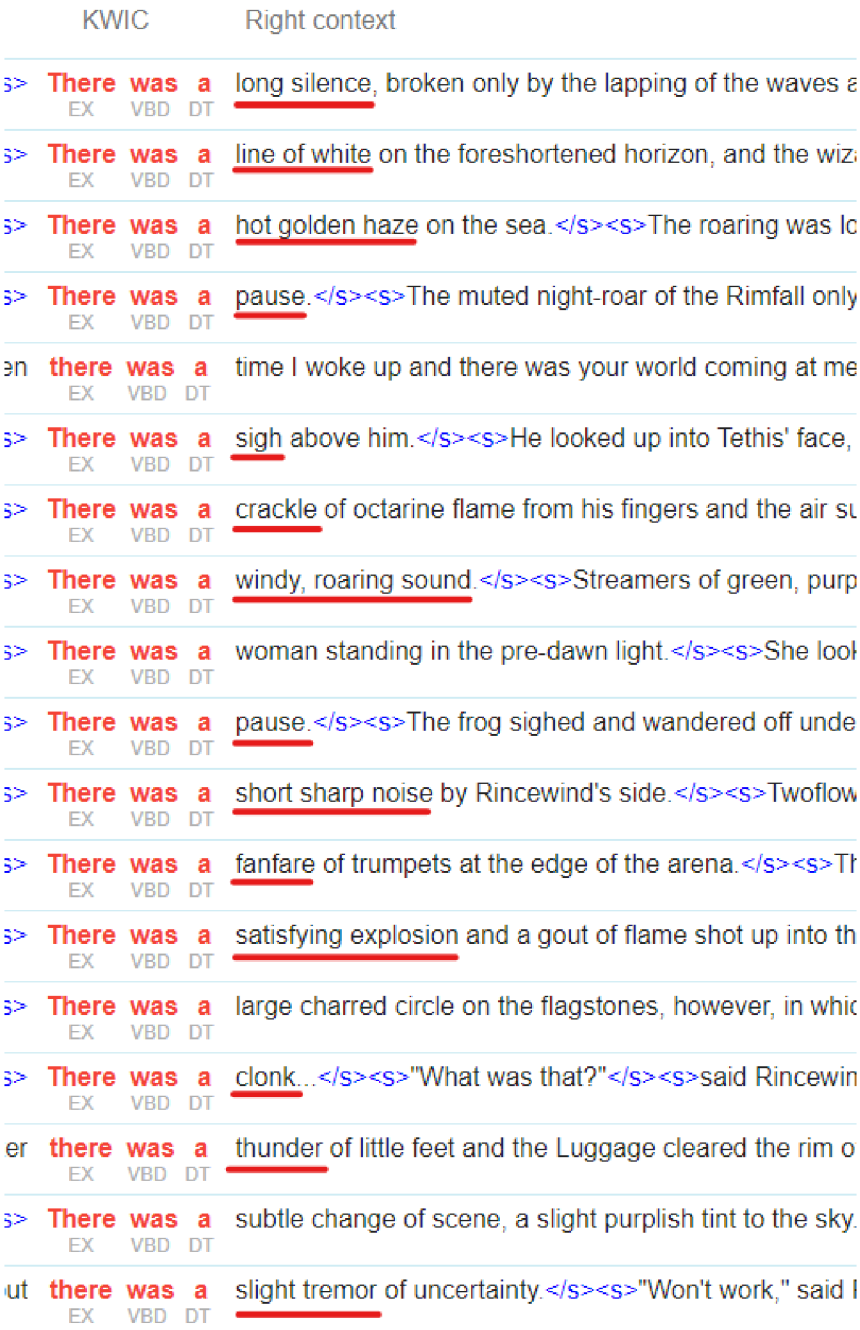

Figure 14: Examples of lemmas that follow the 3-gram "there be a"

When creating a list of 5-grams, it has been observed that the raw frequencies of several of the 10 most common items might be considered too low to be taken into consideration when forming conclusions. For this reason, only the first two clusters have been investigated further as their numbers of occurrences are almost a double of the other 8 items.

Table 19: 5-grams of Colour of Magic (lemmas)

Lemma	Raw frequency
In the centre of the	8
The edge of the world	7
He try not to think	4
There be the faint of	4
On the other side of	4
Try not to think about	4
Find himself look up into	4
Over the edge of the	4
In front of his eye	4
I do not want to	4
Not to think about it	4
Do not appear to be	4
The other side of the	4

The edge of the world would correspond to the aforementioned *over the edge* and *the edge of*. It is another predictable occurrence, as it refers to the setting of the majority of the final chapter (Close to the Edge). *In the centre of the* then usually refers to various locations, such as arenas or rooms (see examples 50 and 51). Therefore, it is also used to write descriptions, which could be considered similar to the *There be not* 3-gram, which was also used for a similar purpose.

(50) *In the centre of the richly decorated room, on a carpet that was so deep and furry that Rincewind trod on it gingerly lest it be some kind of shaggy, floor-loving beast, was a long gleaming table laden with food.*

(51) *There was room beyond them for a rabble of servants and slaves and others who scratched a living here on the roof of the world, and they were all watching the figures clustered in the centre of the grassy arena.*

In conclusion, the N-grams that contain a certain pattern in their usage are often predictable due to the nature of the work of art. Those whose presence cannot be so easily predicted then correspond to discoveries discussed in earlier chapters, such as the frequent dialogues, and could therefore be considered additional evidence.

As mentioned in chapter 2.5, N-grams only take frequency into account, working differently than collocations and colligations, therefore, the amount of new input they bring into an analysis

is rather small (Gray, 2016). This chapter proves that. Although several different n-grams were investigated, they only served to further prove what was already stated in the chapter 4.1 and its subchapters.

4.3. Lexical density

Chapter 2.6 explained the calculation and importance of type-token distribution and its relation to the vocabulary used in the text in question. From this basic description then stems the question: Is Colour of Magic written in a simple language with lexical expressions and phrases often repeated, or is the used language rich, containing a larger number of types when compared to the books of the reference corpus? In order to find the answer, lexical density has to be calculated. The issue with such calculation, however, is the size of the corpus. As stated in the aforementioned chapter 2.6, the larger the corpus, the lower the lexical density due to the natural tendency of English language to feature repeated tokens, specifically of the grammatical category, such as prepositions, articles, conjunctions or pronouns. Therefore, in order to see if Pratchett's vocabulary in the context of Colour of Magic was either complex or simple, only certain word classes will be used for the type-token distribution calculation. The expressions analysed are all going to be of the lexical category, therefore: nouns, adjectives, verbs, and adverbs. Additionally, since the book frequently repeats proper nouns, whose presence does not help prove or disprove the richness of the author's lexis, these tokens and types have been filtered out. An extra category added to the table were then the -ly adverbs, which mainly modify verbs and were therefore investigated mainly to observe in how many unique ways does Pratchett describe the various actions of the characters in Colour of Magic.

The following table lists the results for each category:

Table 20: Type token distribution of lexical word classes in Colour of Magic corpus

Word class	N° of	N° of	Lexical density	Percentage
Nouns	2 780	12 607	0.22	22%
Adjectives	1 291	4 497	0.29	29%
Verbs	1 560	13 360	0.12	12%
Adverbs	602	4 686	0.13	13%
-ly adverbs	407	1252	0.33	33%

From the percentage, it can be seen that Pratchett preferred to use many different adjectives and adverbs, whose lexical densities are always higher than the word class they modify.

To see whether this phenomenon was something specific for Pratchett, a similar table has been constructed based on the 80s fantasy literature corpus (see table 23).

Table 21: Type token distribution of lexical word classes in 80s fantasy literature corpus

Word class	N° of	N° of	Lexical density	Percentage
Nouns	8 112	117 473	0.06	6%
Adjectives	3 892	39 955	0.1	10%
Verbs	3 502	148 561	0.02	2%
Adverbs	1 312	50 881	0.03	3%
-ly adverbs	926	8 503	0.11	11%

Table 21 shows that the phenomena observed in the TTR of different word classes in Colour of Magic corpus are mostly the same as in the case of the 80s fantasy literature corpus. The authors also use more adjectives and adverbs than the nouns and verbs, which are modified by them. As mentioned in chapter 2.6, larger amounts of text naturally result in a lower lexical density, therefore, the results seen in table 21 were partially expectable.

4.4. Keywords

When constructing the lists of keywords, the results obtained from the software had to be closely examined and only certain words were accepted as parts of the final lists of keywords. The most prevailing obstacle was the abundance of proper nouns, which all had to be filtered out, since it is only natural for them to be among the keywords. However, in the case of Terry Pratchett, every noun on the list had to be inspected due to the fact that some of them were used as proper nouns, even though they did not appear so on the first glance. For example, the word *broken* has appeared on the list, but when examined in context, it was apparent that in vast majority of cases, it was used as a part of a name of a pub called Broken Drum. If a word was used for such a purpose, it was not considered as a part of the keywords, just like other proper nouns.

The original idea was to analyse only the keywords of the Colour of Magic corpus, using the 80s Fantasy Literature corpus as a reference. However, in order to see, for example, which words Pratchett tended to avoid, the analysis was also performed the other way, investigating the keywords of the 80s Literature Corpus when using Colour of Magic as reference (see Table 23). This provided another set of data that showed Pratchett's certain habits, which would most likely otherwise be overlooked.

Table 22: Keywords of Colour of Magic corpus, proper nouns filtered out, frequency per million

Keywords – Colour of Magic	Relative frequency
disc	876.23
troll	851.19
guild	638.39
barbarian	275.39
imp	225.32
dryad	212.80
tentacle	200.28
no-one	200.28
arena	200.28
university	162.73
guild	137.69
turtle	137.69
mid-air	125.18
inn-sewer-ant	87.62
materialise	75.11
circumference	75.11
mage	62.59
lizard	62.59
grimoire	62.59
cargo	62.59

Table 23: Keywords of 80s Fantasy Literature corpus, proper nouns filtered out, frequency per million

Keywords – 80s Fantasy Literature	Relative Frequency
gunslinger	654.21
castle	589.75
mother	399.93
uncle	274.58
toward	273.39
gun	229.21
wagon	225.63
dog	188.62
camp	187.43
gray	179.07
honor	159.97
plan	159.97
warband	152.81
guess	142.06
nobody	139.68

north	137.29
darling	121.77
somebody	119.38
chapter	119.38
sick	112.22

4.4.1. Keyword lists findings

Despite the importance of a keyword analysis, the results obtained for this paper did not provide significant insight into the author's idiolect. Since the thesis has already reached a slightly higher word count than expected, a decision has been made to omit mentioning the findings of a keyword analysis in greater detail. The received data did not lead to the formation of new research questions or theories, but rather already confirmed what was mentioned in the earlier chapters. Therefore, the reason for the aforementioned omitting was to avoid unnecessary repetition.

To summarise, the findings led to the discovery that the reference corpus was constructed with a significant flaw. The books comprising the 80s fantasy literature corpus were chosen without taking into account the difference between Englishes (British and American). Therefore, a more thought-out reference corpus would likely bring better results.

Not taking the different variations of English into account, the keyword lists then contain mainly lemmas that are expectable for the books, such as various professions of the main characters, different mythical creatures, or locations where the story takes place and words associated with them (such as *circumference* in the case of Pratchett's Discworld).

5. Conclusion

The main objective of this work was to perform a corpus-based analysis of the book Colour of Magic and try to identify specific features of the author's idiolect. This has been done through two major parts of the practical research. The first was the word list investigation, where the goal was to look into the most frequently used lemmas, find explanation regarding their recurring usage, and among these explanations discover the author's writing habits. The second part of the practical research focused on keyword analysis, using a custom-made reference corpus of popular fantasy literature released in the same decade. The two corpora have also been used both as primary and reference corpora in order to identify not only which words have been mostly exclusive to Pratchett's work, but also to discover which words have been preferred by the other authors while not being used as frequently by Pratchett.

In chapter 1.1.1, a primary research question was asked. Pratchett's writing was said to contain numerous cases of parody and satire; however, such examples could not be seen in the data collected via corpus analysis. Therefore, the answer to the first research question of this paper is no. It is not possible to observe cases of parody and satire in the Colour of Magic only from the data obtained from the focus corpus via the elementary corpus linguistic tools described in this thesis.

When it comes to the parts of the author's idiolect that were observed throughout the paper, it appears that the author had a tendency to avoid including female characters into his work, judging by the near-complete absence of them in Colour of Magic. However, to further prove this, his other works would have to be investigated as well. It is possible that his reason for this stemmed from an easier association with characters of the same sex. There might also be a chance that this phenomenon was progressively disappearing since female characters such as various witches appear in his later-published works. Neither of these theories are proven, however, and could be a topic of another similar research in order to be definitively answered.

Other phenomena discovered during this research are related to Pratchett's approach to direct speech. The lemma *say*, for example, is used rather extensively throughout the book. In the case of both corpora, the lexical expression is present among the most frequent words, however, when the relative frequencies are compared, it is observable that Pratchett uses *say* more often than the other authors. Therefore, it might be safe to say that his novel relies a lot more on dialogue. *Say.* however, has not been the only observed phenomenon related to direct speech. The author also uses various similar lemmas, which differ slightly in meaning to fit the personality of a character. This can be observed with the protagonists, for whom the author used words such as *muttered* or *snapped* as well, to show their particular personalities.

Word choice has not been the only feature observed, however. In the book, there have been discovered several instances of the Pratchett utilizing specific word patterns. This included, for example *as + the dragon + verb*, or *somebody + verb related to sight + with interest*. Although it is impossible to currently explain the author's fondness for such phrases, the numerous examples listed in chapters 4.1.6 and 4.1.7 should not be ignored. In order to determine whether or not such habits were temporary and unique to this book, further research would have to be carried out, looking into the existence of similar phenomena among the other literary works by Terry Pratchett.

In the second part of the research, keywords were investigated, which, contrary to expectations, did not provide much insight into the author's idiolect. The lemmas in the keywords lists were often specific for the different works of art, referring to the professions of the various characters, to species of animals and mythical creatures inhabiting the fantastical worlds, or to the settings of the stories. Therefore, the information obtained through keyword analysis was mostly omitted from the paper to avoid making it unnecessarily lengthy. The only significant discovery of the keyword analysis was that the reference corpus was constructed without taking into account the difference between British and American English. This flaw then led to the finding of various keywords that would otherwise most likely be identical should both the primary and reference corpora work with the same variation of English.

The length of the paper, unfortunately, exceeds the department-established maximum of pages, however, it was impossible to remove any further content due to the rather significant findings and information it may provide. Therefore, it is necessary to state that the thesis ended up being longer than originally anticipated.

Overall, this paper features numerous discoveries related to the Colour of Magic. These do offer certain insight into the author's writing style; however, decisive conclusions should not be formed based on the provided evidence alone. Other research focusing on the specific phenomena described in this work would have to be carried out in order to discover whether or not they are connected only to this particular work of art, or the author's style as a whole.

6. References

- Baker, P. and Egbert, J. 2016. *Triangulating Methodological Approaches in Corpus-Linguistic Research*. London: Routledge.
- Baker, P. and Egbert, J. 2020, *Using Corpus Methods to Triangulate Linguistic Analysis.* London: Routledge.
- Biber, D. and Reppen, R. 2015. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge university press.
- Bondi, M. and Scott, M. 2010. *Keyness in texts - Studies in Corpus Linguistics*, Amsterdam: John Benjamins Publishing Company
- Bowker, L. 2018. "Corpus Linguistics is Not just for Linguists: Considering the Potential of Computer-Based Corpus Methods for Library and Information Science Research." In *Library HiTech* 36, no. 2: 358-371, https://www.proquest.com/scholarly-journals/corpus-linguistics-is-not-just-linguists/docview/2036138529/se-2?accountid=17116 (accessed June 6, 2021).
- Brezina, V., McEnery, T. and Wattam, S. 2015. "Collocations in context – A new perspective on collocation networks." In *International journal of corpus linguistics* 20:2, 139–173. Amsterdam: John Benjamins Publishing Company
- Brezina, V. 2018. *Statistics in Corpus Linguistics.* Cambridge: Cambridge university press.
- Britton, S. 2018. *Thoughtful Laughter: Fantasy and Satire as Social Commentary in Terry Pratchett*, Asheville: University of North Carolina
- Broeder, L. 2007. *Translating Humour - The Problems of Translating Terry Pratchett,* Utrecht: Utrecht University
- Chandler. D, and Munday, R. 2011. "Grammatical words (function words) in *A Dictionary of Media and Communication*, Oxford University Press
- *Collins Online Dictionary,* accessed August 29, 2022, https://www.collinsdictionary.com/dictionary/english/hapax?__cf_chl_tk=yKckWaQ1Sc9nap5Je4XzS4kthl3_h00ASWaQsmK3EbI-1661772898-0-gaNycGzNDT0
- Culpeper, J. 2009. "Keyness. Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet." In *International Journal of Corpus Linguistics* 14:1, pp 29-59. Amsterdam: John Benjamins Publishing Company.

- Gabrielatos, C. 2018. “Keyness analysis: nature, metrics and techniques.” In Taylor, C. & Marchi, A. (eds.) *Corpus Approaches to Discourse: A critical review*. London. Routledge. 225-258.
- Gray, B. 2016. “Lexical bundles.” In Baker, P. and Egbert, J. *Triangulating Methodological Approaches in Corpus-Linguistic Research*. London: Routledge. 13
- Gries, S. Th. 2013. ”50-something years of work on collocations: What is or should be next.“ *International Journal of Corpus Linguistics*, 18(1), 137–166. Amsterdam: John Benjamins Publishing Company
- Hoey, M., Mahlberg M., Stubbs, M. and Teubert, W. 2007. *Text, Discourse and Corpora*, London: Continuum.
- Hyland, K. 2008. “As can be seen: Lexical bundles and disciplinary variation. “In *English for Specific Purposes* 27(1), pp. 4-21
- Kilgarriff, A. and Rychlý, P. 2021. *Sketch Engine User Guide.* Sketch Engine. accessed June 6, https://www.sketchengine.eu/guide/
- Kilgarriff, A. 2009. "Simple maths for keywords." In *Proceedings of the Corpus Linguistics Conference*. Liverpool, UK. https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf
- Lehecka, T. 2015. *Handbook of Pragmatics*, Amsterdam: John Benjamins Publishing Company
- Lindquist, H. 2009. *Corpus Linguistics and the Description of English.* UK: Edinburgh University Press
- Lipka, L. 1992. *An Outline of English Lexicology: Lexical Structure, Word Semantics and Word-Formation*, Tübingen: Niemeyer.
- Louwerse, M. 2004. “Semantic Variation in Idiolect and Sociolect: Corpus Linguistic Evidence from Literary Texts.” In *Computers and the Humanities* 38, 207–221. accessed September 3, https://doi.org/10.1023/B:CHUM.0000031185.88395.b1
- Luthi, D. 2014. "Toying with fantasy: the postmodern playground of Terry Pratchett's Discworld novels." *Mythlore: A Journal of JRR Tolkien, CS Lewis, Charles Williams, and Mythopoeic Literature* 33, no. 1: 8. https://dc.swosu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1036&context=mythlore

- Mahlberg, M. 2009 *Lexical cohesion and corpus linguistics*. Amsterdam: John Benjamins Publishing Company
- *Oxford English Dictionary,* s.v. "idiolect," last accessed 6. 12. 2022, https://www.oxfordlearnersdictionaries.com/definition/english/idiolect?q=idiolect
- Pratchett, T. 1993. *Men at arms*, London: Victor Gollancz Ltd.
- Pratchett, T. 1996. *Maskarade*, London: Corgi
- Scott, M. 2009. "In search of a bad reference corpus." In *What's in a Word-List?* edited by Dawn Archer, 79–91. London: Routledge
- Scott, M. and Tribble, C. 2006. *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins Publishing Company
- Ticak, M. 2021. "No-one, Noone, or No One—Which Is Right?," *Grammarly blog, https://www.grammarly.com/blog/no-one-noone/*
- Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing Company.
- Williamson, G. 2009. "Type-Token Ratio (TTR)," SLTinfo, accessed October 31, https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf
- Xiao, R. and McEnery, T. 2005. "Two approaches to genre analysis: Three genres in modern American English." In *Journal of English Linguistics* 33: 62–82. California: SAGE Publications
- Yarowsky, D. 1993. *One sense per collocation*. Pennsylvania univ. Philadelphia dept. of computer and informational science.
- Yarowsky, D. 1993. "One Sense per Collocation." In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro*, Pennsylvania: University of Pennsylvania, accessed June 4, https://aclanthology.org/H93-1052.pdf