Masaryk University
Faculty of Arts

Department of English and American Studies

English Language and Literature

Eva Václavíková

Idioms of Colour – A Corpus-based Study
Master’s Diploma Thesis

Supervisor: PhDr. Jarmila Fictumová

2010
I declare that I have worked on this thesis independently, using only the primary and secondary sources listed in the bibliography.

.............................................

Author’s signature
I would like to thank my supervisor PhDr. Jarmila Fictumová for her valuable advice and guidance, and particularly for her patience and benevolence.
1 INTRODUCTION ......................................................................................................................... 1

LIST OF ABBREVIATIONS ............................................................................................................. 3

2 THEORETICAL BACKGROUND ................................................................................................. 4

2.1 What is an Idiom? ................................................................................................................... 4
   2.1.1 General description and characteristic features ................................................................. 4
   2.1.2 Classification of Idioms ..................................................................................................... 7
   2.1.3 Origins and development .................................................................................................. 10
   2.1.4 Function of idioms in mass media ...................................................................................... 10

2.2 What is a Corpus? .................................................................................................................. 11
   2.2.1 Corpora in general ............................................................................................................. 11
   2.2.2 Corpus of Contemporary American English ........................................................................ 12
   2.2.3 TIME Magazine Corpus of American English .................................................................... 13
   2.2.4 BYU-BNC: British National Corpus .................................................................................. 14
       2.2.4.1 Additional note ........................................................................................................... 15

2.3 Data Collection Process ....................................................................................................... 16
   2.3.1 Choice of colour terms ...................................................................................................... 16
   2.3.2 Process of collection and narrowing the lists .................................................................... 18

2.4 Colours in Language ................................................................................................................ 19
   2.4.1 Colours and emotions ........................................................................................................ 19
   2.4.2 Semiotics of colours .......................................................................................................... 20

3 ANALYSIS ................................................................................................................................... 23

3.1 General Notes ........................................................................................................................ 23
   3.1.1 Technical considerations .................................................................................................... 23
   3.1.2 Idiomatic classes .............................................................................................................. 25
   3.1.3 Distribution of analyzed phrases in the corpora ................................................................ 26

3.2 Black ...................................................................................................................................... 27
   3.2.1 Frequencies in sections ....................................................................................................... 28
1 INTRODUCTION

“Idioms have often been associated with conversation and informal language, however, the evidence (…) suggests that they are also very common in journalism and magazines, where writers are seeking to write their articles and stories more vivid, interesting to their readers” (Arya)

Languages are living organisms that evolve and change constantly as a result of the need to accommodate new situations in the real world. New words and phrases keep coming to life, while old and/or not frequently used ones are sooner or later forgotten. This process seems to be most obvious in idiomatic expressions. “Catchphrases, fads, gimmicks hold the popular fancy for a short while and then disappear. Others meet a particular need and survive whilst they have a function to fulfil” (Flavell 19). But as soon as there is no longer a situation, event or condition it could describe, the idiom is almost certainly doomed.

Native speakers use idioms in every day communication (Correli). Spoken language is undoubtedly closest to the speaker, as it is the speaker who creates the words, sentences and utterances, and this is where new expressions are brought to life. Then there is the language used in newspapers and magazines; texts in these media are created usually on a daily or weekly basis and thus record the language in its current use. Newspapers also help to spread new expressions, such as neologisms, freshly coined idioms and catchphrases, because “they are the most widely, daily read mass media of written communication” (Arya). On the other side of the scale is the language of academic publications (and fiction to a certain extent), because these are more or less formal works, where the rather informal idioms are disfavoured and special, technical and scholarly terms are used instead; moreover the creation of such works can take
months or even years and thus the temporary fads and trends may not survive long enough or be popular enough to appear in them.

This thesis aims to look at the group of colour idioms in American English and to analyze their frequency in different sections of language or genres. I chose the semantic field of colours, because colours are everywhere around us and they are very frequently used to convey certain meanings (for example, on traffic signs and traffic lights); therefore, colours are bound to appear in everyday language, not excepting idiomatic phrases. I assume that the analyzed expressions are more frequent in newspapers and magazines than in other types of texts, academic texts in particular. For the purpose of verification of this hypothesis online corpora are used.

The theoretical part of this thesis provides background information on idioms: some general definitions and characteristic features are mentioned and explained, one subsection is devoted to the origins of idiomatic expressions, and a classification of idioms is indicated, as well as a brief description of their purpose in language. The second part of Chapter 2 deals with online corpora: besides general information about the three text corpora used in the analysis, the interface is briefly explained as well. Section 2.4 is devoted to colour semiotics. What follows is the analysis of the collected data, supplemented with notes on the results of my research and some interesting facts related to the changes in popularity of several specifically chosen idioms over time.
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acad.</td>
<td>Academic</td>
</tr>
<tr>
<td>BiWec</td>
<td>Big Web Corpus</td>
</tr>
<tr>
<td>BNC (BYU-BNC)</td>
<td>British National Corpus</td>
</tr>
<tr>
<td>COCA</td>
<td>Corpus of Contemporary American English</td>
</tr>
<tr>
<td>Fict.</td>
<td>Fiction</td>
</tr>
<tr>
<td>Mag.</td>
<td>Magazine</td>
</tr>
<tr>
<td>Misc.</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>News.</td>
<td>Newspaper</td>
</tr>
<tr>
<td>NMC</td>
<td>New Model Corpus</td>
</tr>
<tr>
<td>Sp.</td>
<td>Spoken</td>
</tr>
</tbody>
</table>
2 THEORETICAL BACKGROUND

2.1 What is an Idiom?

2.1.1 General description and characteristic features

What is an idiom? That is a rather tricky question, because there is no single definition of idiom and there are no hard and fast rules that would conveniently and unexceptionably describe it. “[I]n general, idiomatic expressions are very versatile. This is the main reason for numerous difficulties with defining precise boundaries of idioms” (Tkachuk 71). The boundary between idioms and collocations is not clear. Tkachuk says that “the constituents of collocations retain their original meaning, while the constituents of idioms acquire a new meaning as a whole” (64). However, she admits that although the two criteria suffice for the majority of idioms, marginal cases are more difficult to deal with and they can be moved to either of the two categories.

It can be said that idioms are ready-made phrases and commonplace expressions natural to native speakers. In most publications on idioms, the definitions mention the peculiar behaviour of idioms as far as meaning is concerned. Macmillan dictionary tells us that idiom is “an expression whose meaning is different from the meaning of the individual words” (710). Correli defines idiom as an “expression that does not mean what it literally says”; an expression whose “meaning is often quite different from the word-for-word translation”. Similarly, Flavells say that “idioms are mostly phrases that can have a literal meaning in one context but a totally different sense in another” and they elaborate further on this idea of hidden meaning, stating that:

the words in [idioms] do not mean what they ought to mean – an idiom cannot be understood literally. (…) The meaning of the whole, then, is
not the sum of the meaning of the parts, but is something apparently quite unconnected to them. (6)

Malá defines phrasal idioms as “stable and rooted phrases, whose fundamental characteristic features are *polylexicality* (they consist of at least two lexemes) and *stability*”¹. Here, it is necessary to note that polylexicality, or lexical complexity in other words, is a feature questioned by several linguists. On page 64 of her paper, Tkachuk mentions Makkai, who allows also for one-word idioms. Moreover, if an idiom had to consist of more than one word to be classified as an idiom, where would verbs such as “blackleg”, “blackball”, and “blackmail” belong? Besides polylexicality and relative structural stability, Malá also mentions *idiomaticity* as one characteristic. “Idiomaticity is a property of a phrase, which means that the meaning of the whole phrase cannot be deduced from the meaning of individual words”². Based on that, she describes idioms as those expressions that have a certain degree of idiomaticity; while expressions without idiomaticity or with a very low degree of idiomaticity are labelled as collocations.

Langlotz describes idiomatic constructions as “complex symbols with specific formal, semantic, pragmatic and sociolinguistic characteristics” (3).

---

¹ “Jako frazeologizmy/frazémy jsou pojímána pevná (stabilní) a zaužívaná slovní spojení, jejichž základními znaky jsou polylexikalita (sestávají minimálně ze dvou lexémů) a stabilita.” – Malá

² “Idiomaticitu se rozumí ona vlastnost slovního spojení, kdy se význam celého spojení nedá odvodit z významů jednotlivých slov (…)” – Malá
Institutionalisation refers to “an idiom’s degree of familiarity within a speech community”. The term compositeness “refers to the fact that idioms are multi-word units that consist of two or more lexical constituents”. Frozenness or fixedness (another term adopted by Langlotz) means that the phrase is lexically and grammatically restricted. The last term, non-compositionality, “points to the fact that the meaning of these constructions is not derivational sum of the meaning of their constituents” (3-4). This lastly mentioned feature of idioms is traditionally used as the primary feature to define them – see for example the definitions of idiom above, at the beginning of this section.

Similarly, in her work on money idioms, Tkachuk summarizes the following features of idioms found in most works and encyclopaedias (63):

- lexical complexity and semantic simplicity
- fixed form or invariability
- figurativeness

By lexical complexity and semantic simplicity she means that “the meaning of an idiom is not derivable from the meaning of its individual parts” (63). Invariability is in fact another name for Langlotz’s fixedness or frozenness: “the number of paradigmatic
variants [of an idiom] (if there are any) is limited” (63). However, a few sections later, she adds that this criterion is true only to a certain extent, because idioms have variants - they can be syntagmatic, paradigmatic, grammatical, spelling (British “grey area” versus American “gray area”), or geographical (British “have green fingers” versus American “have a green thumb”) (65). The criterion of figurativeness is also questioned by Tkatchuk. She claims that “not all idioms are equally figurative”. She divides idiomatic phrases into phraseological fusions, which are “absolutely non-motivated idioms, at least from the perspective of modern English”, phraseological unities, which are partially motivated, and phraseological combinations (i.e. motivated idioms). “Motivated idioms have a greater potential to express the semantics of idioms through the semantics of individual words” (68); an example of a motivated idiom could be the phrase a bolt from the blue, which clearly refers to the unexpected appearance of a bolt of lightning in the clear sky.

2.1.2 Classification of Idioms

There are several possibilities, how to categorize and classify idioms. Granger and Meunier attribute “one of the most influential typologies in English lexicology and lexicography” to Cowie, who “makes a primary distinction between composites, which function syntactically at or below the level of the sentence, and formulae, which function pragmatically as autonomous utterances” (64).
At the beginning of his dictionary of idioms, Cowie says that “there is no clear dividing-line between idioms and non-idioms: they form the end-points of a continuum” (xiv). Nevertheless, he categorizes idiomatic expressions according to their degree of fixedness and figurativeness. He distinguishes between:

- Pure idioms
- Figurative idioms
- Restricted collocations
- Open collocations

**Pure idioms** “form the end-point of a process by which word-combinations first establish themselves through constant re-use, then undergo figurative extension and finally petrify or congeal”. In other words, pure idioms are those expressions to which Langlotz’s and Tkachuk’s features mentioned in the previous section can be applied without exception. In **figurative idioms**, “variation is seldom found (…) and pronoun substitution unlikely”. As far as the transparency of meaning is concerned, “the literal senses of these expressions do not survive alongside their figurative ones in normal, everyday use”. **Restricted collocations** are sometimes called “semi-idioms”. In these phrases, “one word (…) has a figurative sense not found outside that limited context. The other element appears in a familiar, literal sense”. Some expressions in this category allow a degree of lexical variation. In **open collocations**, the elements are freely recombinable and each element is typically used in a common literal sense.

(Cowie xiv)

In *Understanding figurative language*, Glucksberg (qtd. in Arya) identified four types of idioms:
• non-compositional/opaque (e.g. by and large)
• compositional/opaque (e.g. kick the bucket)
• compositional/transparent (e.g. spill the beans)
• quasi-metaphorical (e.g. don’t give up the ship)

The term non-compositional means that idioms cannot be analyzed either semantically or syntactically, while compositional idioms can. The term opaque refers to the fact that meanings cannot be derived from the meanings of the constituents; meanings of transparent idioms, on the other hand, can be mapped onto their constituent words.

Idioms of the fourth type, quasi-metaphorical, are idioms, “which behave just as do metaphors” (sic). In a way, Glucksberg’s categorization is similar to Cowie’s in that the categories form a continuum going from petrified, opaque expressions to those that allow some variation and whose meaning can be arrived at through semantic analysis.

Tkachuk (65) provides the following division into idiomatic classes, based on grammatical characteristics of the phrases:

• sentential (e.g. The grass is always greener on the other side of the fence.)
• non-sentential
  • verbal (e.g. blacken sb’s reputation)*
  • verbless
    • nominal (e.g. a red rag to a bull)
    • adjectival (e.g. white as snow)
    • adverbial (e.g. once in a blue moon)

* Purely verbal idioms consist of verbs only, but they are not very frequent. The majority of so-called verbal idioms comprise a predicate and an object, as above. (65)

Tkatchuk’s classification is the one used for categorization of phrases in the practical part of this thesis.
2.1.3 Origins and development

Tkachuk states that “idioms are inclined to informal usage”, but their stylistic features “are tightly interconnected with their origins” (70). She found that “idioms of Biblical origin are more frequent in written language, while those originating from slang expressions or gambling usually belong to informal speech” (70).

Literature is an excellent source of idioms. Besides the Bible, Shakespeare has been another very potent source of idioms: The Phrase Finder lists 135 phrases coined by William Shakespeare (and 70 phrases from the Bible). Among other literary works that brought new expressions into life or popularized less frequently used ones is definitely Carroll’s Alice in Wonderland. New idioms are also formed as a result of memorable events, but Flavells note that “there is a high drop-out rate” (159). Popular culture enriches the vocabulary, for example via catch phrases.

As Flavells in their Introduction to Dictionary of Idioms and their Origins point out, it is “extremely difficult to state accurately when the idiom was first used – as an idiom” (vi). That is because “phrases have literal meanings, then they generally develop metaphorical uses and ultimately, in typical cases, acquire an idiomatic sense that is separate from the literal one”. What is now considered to be a pure idiom could have been a perfectly understandable expression with a very different meaning only a few years or decades ago.

2.1.4 Function of idioms in mass media

Malá, in her conference contribution on stylistic use of idioms in mass media, summarizes the functions of phrases. She says that “[t]he use of an idiom is more convincing, more effective; it awakens an interest in the recipient, because the expression may be humorous, ironic. Idioms are, therefore, often means of humour and
Phrasal idioms are much more expressive and make the utterances and evaluations more intense. Malá further adds that:

> The function of idioms lies not only in the intensification and augmentation of expressivity/emotionality of the text, but they also have a function in text structuralization. Idioms serve as synonymic variants of lexemes and they allow for textual progression.

### 2.2 What is a corpus?

#### 2.2.1 Corpora in general

Simply put, a corpus is a body of texts; it is a collection of texts stored on a computer and selected according to specific criteria. The Macmillan dictionary defines corpus as “a collection of written and spoken language stored on computer and used for language research and writing dictionaries” (312). Text corpora have generally one common role: “exploration of variation in language, according to, for example, genre, domain or region, or over time” (Pomikálek 2). Online corpora, such as those discussed below, allow users to search for words and phrases by exact word or phrase (e.g. *black sheep*), wildcard or part of speech (e.g. *white of [p]*, where [p] stands for pronouns), or combinations of these (e.g. *until * [be] blue in the face*, which combines a wildcard * with lemmatized verb). It is also possible to search for collocates and find the frequency of words and phrases in any combination of registers. In addition, registers can be

---

3 Použití idiomu působí přesvědčivěji, efektivněji, vzbudí větší pozornost recipienta, protože může být vtipně, ironické. Idiomy jsou tedy nezřídka prostředkem humoru a satiry. – Malá

4 Idiomy nemají jenom funkci intenzifikace a zvyšování expresivity/emocionality textu, nýbrž vykonávají i funkci při strukturalizaci textu. Slouží jako synonymní varianty k lexémům a zajišťují textovou progresi (...) – Malá
compared and semantically-based queries of the corpus are easily carried out. Besides, corpora generate lists of words used more in one period than another; they help to show how the meaning of words has changed over time, by looking at changes in collocates.

Among the best-known online corpora is the BYU-BNC: British National Corpus (hereafter BNC) and the Corpus of Contemporary American English (hereafter COCA), both created by the professor of corpus linguistics at Brigham Young University in Utah, Mark Davies. Davies created also another useful tool, the TIME Magazine Corpus of American English (hereafter TIME). These three corpora are available free of charge to wide public. I will be working with them in the practical part of this paper.

<table>
<thead>
<tr>
<th>Corpus of Contemporary American English</th>
<th>400+ million words (US 1990-present)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME Magazine Corpus</td>
<td>100 million words (US 1923-2006)</td>
</tr>
<tr>
<td>BYU-BNC: British National Corpus</td>
<td>100 million words (UK 1980s-1993)</td>
</tr>
</tbody>
</table>

Table 1: Overview of numbers of words and periods covered by the three used corpora

2.2.2 Corpus of Contemporary American English

The COCA corpus is:

the largest freely-available corpus of English, and the only large and balanced corpus of American English. (…) Because of its design, it is perhaps the only corpus of English that is suitable for looking at current, ongoing changes in the language. (Davies, COCA)

The COCA consists of texts from five different genres: spoken language, fiction, popular magazines, newspapers, and academic journals; these categories make up for approximately even parts of the corpus: each containing from 79 to 84 million words, which gives the total of 400 million words. The exact numbers (in millions) are:

- Spoken – 81.7
The corpus is highly representative of contemporary American English, as it is an open corpus that is updated on regular basis once or twice a year, and each year contains about 20 million words (Davies, *Corpus*).

The interface of the corpus is very user-friendly; still, it makes complex searches possible and easily done. The results can be displayed as a list or in a chart, divided into sections by genre or by years. The material can be searched simply by words (*blackball*), phrases (*red tape*), lemmas (*blackout*), and using wildcards “?” and “*”, but more complex searches are possible as well via tags for parts of speech (the list of tags is available right in the interface). The results may be clicked, which takes the user to a window with context and further information about the source text. Each occurrence is on a separate line; if clicked, even more detailed source information and an expanded context appear. Using the “Sections” option, the search can be limited to certain genres and subgenres or to certain periods of time, and these sections can be compared. The interface allows the user to order the listing of results as s/he finds fit (by section name, number of occurrences per million, or by number of tokens).

### 2.2.3 TIME Magazine Corpus of American English

The TIME corpus is smaller in number of words, but is more specialized and extends over a larger period of time than the COCA. It allows the user to conveniently search more than 100 million words of texts published in TIME magazine from 1923 to 2006. The user can see “how words, phrases and grammatical constructions have increased or
decreased in frequency and see how words have changed meaning over time” (Davies, *Time*).

The interface is the same as that of the COCA; the only difference being that all the texts in the TIME corpus come from the magazine of the same name, and therefore are divided only to time periods (decades or individual years), not genres. This corpus is used in the practical part of this thesis mostly for reference, as it does not allow comparison between different registers, but shows only frequencies of expressions in magazines.

2.2.4 BYU-BNC: British National Corpus

The BNC corpus is in many respects similar to the above mentioned COCA and TIME corpora. However, it is based solely on British texts. It is a collection of sample texts of the same size as the TIME – 100 million words “of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written” (“What is the BNC?”). The written samples come from various kinds of texts: newspapers, journals, periodicals, fiction, academic essays etc. The spoken part of the corpus, which amounts to 10 % of the words, includes “orthographic transcriptions of unscripted informal conversations (…) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins” (“What is the BNC?”). The words are divided into the sections as follows (in millions):

- Spoken – 10.0
- Fiction – 15.9
- Newspaper – 10.5
- Academic – 15.3
- Miscellaneous – 44.6
The corpus was created between 1991 and 1994. Since then, no new texts have been added. In 2001 and 2007, BNC World and BNC XML Edition respectively were released (with minor revisions only); unfortunately, neither of these is available online to public.

Mark Davie’s BYU-BNC is only one of several online services based on the BNC (others are, for example, BNC Web and Just the Word). This service allows “to search for a word or a phrase in the BNC but will only show a limited amount of context” ("Other services"). The interface is the same as the one described above in section 2.2.1., except for the fact that the texts are not divided into sections by the time of their creation.

Because of the obsoleteness of this corpus and its relatively small size, there have been attempts, recently, to create a bigger and better corpus. Pomikálek et al. claim that “a corpus should be big enough to have subcorpora where the expected frequency for the item being studied is at least thirty or forty” (2). While the BNC may be sufficient for high frequency words, it does not provide enough information to make informed generalisations on phrasal and collocational items (2). Therefore, Pomikálek, Rychlý, and Kilgarriff created the Big Web Corpus (BiWeC), which contains 5.5 billion words so far (the target size is 20 billions). In a parallel stream of work they developed the New Model Corpus: an open, collaborative corpus, freely available for research purposes that aims to replace the BNC (10). These new corpora are created with the use of special software tools, which can encode and query them efficiently; they contain less noise and fewer duplicates, and are much faster.

2.2.4.1 Additional note

Quite unexpectedly, the BYU-BNC corpus was updated at the end of March 2010. Two sections were added – “Magazine” and “Non-academic”. As a result the distribution of
words in section “Miscellaneous” changed, because the two sections were created by extracting texts from it. The numbers of these three sections in the new version are: Magazine – 7.3, Non-academic – 16.5, Miscellaneous – 20.8 (in millions of words). The size of the other sections remained the same. As this change happened only after I collected the necessary data, the analysis in Chapter 3 is based on the old version of the corpus.

2.3 Data collection process

2.3.1 Choice of colour terms

In order to keep the process of data collection organized and transparent, and to legitimate my choice of colour words for the analysis, I first needed to create a list of colours. At first, I tried to work with the list found in the software dictionary Lingea Lexicon 2002, ver. 4.10 (it can be accessed after typing e.g. “red” in the search window - the list itself is to be found in the lower left pane). However, that list is rather extensive and includes expressions that are not colours in the real sense - see table 2 below.

<table>
<thead>
<tr>
<th>amber</th>
<th>ecru</th>
<th>orange</th>
<th>straw-coloured</th>
</tr>
</thead>
<tbody>
<tr>
<td>azure</td>
<td>emerald</td>
<td>pastel</td>
<td>strawberry blonde</td>
</tr>
<tr>
<td>beige</td>
<td>faded</td>
<td>peach</td>
<td>swarthy</td>
</tr>
<tr>
<td>black</td>
<td>fawn</td>
<td>pink</td>
<td>tan</td>
</tr>
<tr>
<td>black and white</td>
<td>florid</td>
<td>pitch-black</td>
<td>ultramarine</td>
</tr>
<tr>
<td>blackish</td>
<td>gamboge</td>
<td>puce</td>
<td>vermilion</td>
</tr>
<tr>
<td>blood-red</td>
<td>gold</td>
<td>purple</td>
<td>violet</td>
</tr>
<tr>
<td>blue</td>
<td>golden</td>
<td>purplish</td>
<td>washed-out</td>
</tr>
<tr>
<td>bluish</td>
<td>green</td>
<td>red</td>
<td>white</td>
</tr>
<tr>
<td>brown</td>
<td>greenish</td>
<td>reddish</td>
<td>yellow</td>
</tr>
<tr>
<td>brownish</td>
<td>greenish blue</td>
<td>reddish brown</td>
<td>yellowish</td>
</tr>
</tbody>
</table>
Although I tried to limit the list methodically by excluding certain groups of words (those that do not describe a colour property by themselves, such as “dark” and “deep”; compound words, such as “blood-red” and “sky-blue”; words ending in –ish, such as “blackish” and “bluish”; and words that in addition to denoting a quality (colour) are also used as nouns to denote an object – e.g. “amber” and “chestnut”), I could not come to a brief list that would include only those words that can be unarguably considered colour words. Therefore, I tried to look at the problem from a more technical point of view – wavelength. I found that Newton was the first to divide the continuous spectrum into seven divisions: Red, Orange, Yellow, Green, Blue, Indigo and Violet (“Colour”). Indigo, a colour between blue and violet, is not usually recognized as a separate division by modern colour scientists, therefore, I excluded it from the list. Nevertheless, I added three greyscale colours: White, Grey, and Black.

<table>
<thead>
<tr>
<th>buff</th>
<th>grey</th>
<th>rose</th>
<th>yellowy</th>
</tr>
</thead>
<tbody>
<tr>
<td>burgundy</td>
<td>greyish</td>
<td>rosy</td>
<td></td>
</tr>
<tr>
<td>carmine</td>
<td>hazel</td>
<td>ruby</td>
<td></td>
</tr>
<tr>
<td>chestnut</td>
<td>indigo</td>
<td>ruddy</td>
<td></td>
</tr>
<tr>
<td>cerise</td>
<td>iron blue</td>
<td>russet</td>
<td></td>
</tr>
<tr>
<td>claret</td>
<td>ivory</td>
<td>saffron</td>
<td></td>
</tr>
<tr>
<td>copper</td>
<td>jade</td>
<td>sallow</td>
<td></td>
</tr>
<tr>
<td>coral</td>
<td>luteous</td>
<td>salmon pink</td>
<td></td>
</tr>
<tr>
<td>crimson</td>
<td>magenta</td>
<td>scarlet</td>
<td></td>
</tr>
<tr>
<td>dark</td>
<td>maroon</td>
<td>sepia</td>
<td></td>
</tr>
<tr>
<td>dark-blue</td>
<td>mauve</td>
<td>silver</td>
<td></td>
</tr>
<tr>
<td>dark-brown</td>
<td>navy</td>
<td>silvery</td>
<td></td>
</tr>
<tr>
<td>deep</td>
<td>ochre</td>
<td>sky-blue</td>
<td></td>
</tr>
<tr>
<td>dun</td>
<td>olive</td>
<td>snow-white</td>
<td></td>
</tr>
</tbody>
</table>

*Table 2: List of colours taken from Lingea Lexicon 2002, ver. 4.10*
Thus, the idiomatic expressions I was looking for were those, which included one or more words from the following list:

- black,
- blue,
- green,
- grey/gray,
- orange,
- red,
- violet,
- white, and
- yellow.

The terms are all used frequently and I, therefore, anticipated that idiomatic expressions would not be difficult to find.

2.3.2 Process of collection and narrowing the lists

I began my search for the relevant expressions in general English dictionaries from well-known and renowned publishers: Macmillan, Collins Cobuild, Oxford and Cambridge. What followed was a search in specialized dictionaries of idioms from various publishers with the purpose of checking the expressions and possibly finding less frequently used ones.

Names of plants and animals were excluded completely from the analysis, because they are not idiomatic and were initially collected only as side-notes. So were the trademarks and brand names (“Yellow Pages”, “Greyhound Bus”, and “Bluetooth”). Expressions denoting special events or situations in the US and UK, which are closely related to the culture (e.g. “Orange Bowl”, and “Orange Prize”) as well as names of
food and meals ("Black Forrest gateau", “black bun” etc.) were not included in the analysis either.

2.4 Colours in language

2.4.1 Colours and emotions

Anders Steinvall conducted a very interesting research on colours and emotions in English. Using the Bank of English corpus, he explored the collocations of 50 colour terms and 135 emotion terms, and investigated “which color categories are salient from the viewpoint of emotions and vice versa” (348). Based on his results, Steinvall states that “English speakers maintain a system of preference by which they link colors and emotions (…)” (350).

<table>
<thead>
<tr>
<th></th>
<th>LOVE</th>
<th>JOY</th>
<th>SURPRISE</th>
<th>ANGER</th>
<th>SADNESS</th>
<th>FEAR</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLACK</td>
<td>0.65</td>
<td>0.65</td>
<td>0.00</td>
<td>16.13</td>
<td>75.48</td>
<td>7.10</td>
<td>100</td>
</tr>
<tr>
<td>WHITE</td>
<td>2.53</td>
<td>18.99</td>
<td>0.00</td>
<td>39.24</td>
<td>5.06</td>
<td>34.18</td>
<td>100</td>
</tr>
<tr>
<td>RED</td>
<td>7.23</td>
<td>12.72</td>
<td>0.00</td>
<td>55.49</td>
<td>19.94</td>
<td>4.62</td>
<td>100</td>
</tr>
<tr>
<td>YELLOW</td>
<td>3.70</td>
<td>81.48</td>
<td>0.00</td>
<td>3.70</td>
<td>7.41</td>
<td>3.70</td>
<td>100</td>
</tr>
<tr>
<td>GREEN</td>
<td>0.50</td>
<td>5.97</td>
<td>0.00</td>
<td>88.06</td>
<td>5.47</td>
<td>0.00</td>
<td>100</td>
</tr>
<tr>
<td>BLUE</td>
<td>0.00</td>
<td>54.90</td>
<td>0.00</td>
<td>3.92</td>
<td>37.25</td>
<td>3.92</td>
<td>100</td>
</tr>
<tr>
<td>BROWN</td>
<td>0.00</td>
<td>7.14</td>
<td>0.00</td>
<td>50.00</td>
<td>42.86</td>
<td>0.00</td>
<td>100</td>
</tr>
<tr>
<td>GREY</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>4.00</td>
<td>75.00</td>
<td>20.00</td>
<td>100</td>
</tr>
<tr>
<td>PINK</td>
<td>30.43</td>
<td>43.48</td>
<td>0.00</td>
<td>11.59</td>
<td>8.70</td>
<td>5.80</td>
<td>100</td>
</tr>
<tr>
<td>ORANGE</td>
<td>8.70</td>
<td>73.91</td>
<td>0.00</td>
<td>13.04</td>
<td>4.35</td>
<td>0.00</td>
<td>100</td>
</tr>
<tr>
<td>PURPLE</td>
<td>12.96</td>
<td>5.56</td>
<td>0.00</td>
<td>57.41</td>
<td>18.52</td>
<td>5.56</td>
<td>100</td>
</tr>
</tbody>
</table>

Image 3: Salience of basic emotion categories in reference to basic colour categories in percentage of instances (Steinvall 357)

The most frequent collocation of black in Steinvall’s sample is sadness. He explains it by the fact that “black is the color of death and mourning in European and American cultures, and this should be reflected in language” (357-8). Grey is also associated with sadness and Steinvall believes that the link is metaphoric of weather.
White is strongly associated with anger and fear, and the connection is motivated by facial colour. Red is represented in many categories, but the most frequent is the collocation with anger. Steinvall suggests that facial colour and heat are the motivation here. Both yellow and orange collocate most with joy. Green is associated with anger category, precisely with envy and jealousy. What Steinvall highlights as striking, is the collocation of blue with joy, not only with sadness. Nevertheless, he suggests, that the results may be distorted here, because blue’s association with melancholy and sadness is so strong, that the word for the emotion is usually completely left out (and, therefore, such instance is not found and recorded by the researcher). (358-9)

2.4.2 Semiotics of colours

Although Vaňková’s work is based on Czech rather than English, her findings are applicable to other languages as well, especially the European ones, because the cultural background is shared to a certain extent and the languages have common origins. She summarizes how people interpret colours:

We always perceive the “colorful” as positive: colorfulness is associated with light, beauty, life, wealth, good luck, joy, diversity of full life. The non-colorful – characterized often as “colorless”, grey, black-and-white; and pale when a human being is concerned – represents the other extreme: gloom, ugliness, a poor spiritual and physical life, the oppressive weight of everyday life, poverty, sadness, boredom, lethargy, disgust, and monotony. (441)

Further in the text, Vaňková clarifies that “the physically non-colourful (achromatic!) colors black, white and grey are in the position of non-colors also in language” (443).

Flavell, on page 30, provides a short essay on the meaning of black and white colour in idioms. He notes that “[i]n some areas of life, in art or the church for instance,
black symbolizes evil” (e.g. as black as the devil, as black as hell). Besides evil and wrong-doing, black colour is also associated with illegality and law-breaking (e.g. black market, black economy, blackmail). “Evil and illegality obviously bring moral censure and disgrace. Not surprisingly then there are plenty of phrases expressing this idea” (e.g. to blackball, a blackleg, the black sheep of the family). Congruently with Steinvall, Flavell also mentions the association of black with death in most cultures and he believes that “this probably explains the gloomy connotations of the word in relation to human feelings” (be in a black humour/mood, paint things in black colours). Some idioms “suggest anger and threat” (e.g. give somebody a black look).

On the other hand, “white has the power to turn something bad into something good. Lying, witches and magic all have negative associations, yet add the positive word white and they are rendered harmless, even beneficial: a white lie, a white witch and white magic.” Although there are some negative expressions, where white colour is used, they are very sparse (e.g. white-livered, show the white feather). Flavell summarizes that “it is generally true to say that in English black indicates bad whilst white indicates good” (30).

Green is the colour of plants and vegetative life. According to Vaňková, “[i]f green is used in connection with a human being, it always has a negative connotation (…). In relation to the human body, green always means illness, even closeness of death and decay.” (448)

Although yellow is a joyful colour, the colour of sun and flowers, it also signalizes change for the worse:

[yellow] is associated not only with old age, but also with illness – usually a more serious and longer illness and even death. Another
important field connected with a complexion changed to yellow are negative emotions – jealousy, envy and hatred. (449)

Vaňková explains this association of yellow with negative emotions (which Steinvall connects rather with the green colour) by the fact, that yellow is etymologically related to green. Moreover, yellow and green are very close to each other on the colour spectrum.

Blue is the most typical cold colour, “color suggesting the sky, air, cold, emotional distance and spirituality”. It can be connected with suffering from cold or lack of air, but it also refers to bruises (to be black and blue). Vaňková summarizes the meaning of blue as follows:

In association with the human body, blue also connotes physical discomfort, and, even more strongly than green, danger to life. (…) In general, we can say that it is an extreme pole of paleness, even an attribute of death (449).
3 Analysis

3.1. General notes

In the preliminary research, I collected around 500 expressions that contained some of the chosen colours. Their degree of idiomaticity varied greatly: some were pure idioms, some were mere collocations. For the analysis I picked only the most interesting ones and I tried to include idioms rather than collocations. However, as there is no definitive boundary between these, my choice was rather subjective, based mainly on my perception of the expression. Still, I tried to follow the classification suggested by various linguists and outlined in chapter 2.1.2. I included metaphorical expressions such as red as a beetroot, white as snow, and green as grass, even though Malá states that such transparent similes belong to the group of collocations, not idioms.

3.1.1 Technical considerations

I went through the filtered list of phrases, checking each phrase in the three online corpora and noting their numbers of occurrences. This was meant to show, whether there are any expressions used solely in American or British variety. However, this task proved to be rather complicated and time-consuming: it was a demanding and difficult job to filter out only the appropriate tokens of the expressions. For example, the search for the phrase (to be) in the black, which means that one has no debts or earns more money than s/he spends, showed a lot of instances of absolutely unrelated expressions. The total was 1876 occurrences in the COCA, but these included sentences such as “we were in the Black Sea”, “He stood at the door in the black jogging pants”, etc., and it was virtually impossible to separate relevant findings from irrelevant ones. Another example of such problematic expression is the phrase yellow-belly or yellow-bellied –
the latter is used frequently in connection with various animals (yellow-bellied sap-sucker and yellow-bellied marmot being the most frequent ones). Having gone through the list of results manually, I found that out of the 33 occurrences (in the COCA) only 8 were related to the idiomatic phrase connected with fear and cowardice. The context of expressions with low numbers of occurrences was inspected one by one and the irrelevant cases were excluded. The popular expressions, which yielded hundreds of tokens in the three corpora, were explored only at random – when not many problematic places were found, the numbers were noted down as they appeared in the corpus.

What also required some thought was how to search for more complicated expressions, in particular those consisting of several words, including pronouns, or taking objects. If not properly entered, such search resulted in irrelevant phrases. In order to receive only the desired expressions, limitations of the corpora’s interface had to be taken into account as well (e.g., the “Context” field accepts only a single word).

For example, the idiomatic phrase to give someone/something the green light could be entered as “[give] * the green light” or in the following way:

![Image 4: Search window in COCA]

The second option, which made use of the “Context” field, made it possible to easily and quickly look up both give someone/something a/the green light and give a/the green light to someone/something in different tenses and aspects, with definite as well as
indefinite articles and with various objects. Differently entered phrase yielded different results – see the comparison of the discussed idiom below.

<table>
<thead>
<tr>
<th></th>
<th>COCA</th>
<th>S.</th>
<th>F.</th>
<th>M.</th>
<th>N.</th>
<th>A.</th>
<th>BNC</th>
<th>S.</th>
<th>F.</th>
<th>N.</th>
<th>A.</th>
<th>Mis.</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>[give] * the green light</td>
<td>63</td>
<td>13</td>
<td>7</td>
<td>27</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>green light + [give] in context</td>
<td>276</td>
<td>90</td>
<td>16</td>
<td>67</td>
<td>83</td>
<td>20</td>
<td>39</td>
<td>-</td>
<td>3</td>
<td>19</td>
<td>1</td>
<td>16</td>
<td>104</td>
</tr>
</tbody>
</table>

*Table 3: Comparison of different inputs of the same expression*

It is possible that my results will differ from those others may find, because there were usually more ways of entering the expression into the interface. I did my best to find only the relevant occurrences of phrases, but there might have been a better option that had not occurred to me. Those phrases that were found in the dictionaries, but were not present in any of the three corpora, were deleted from the list and not dealt with at all. A few phrases had to be skipped, because I was not able to filter out only the relevant data from the results – these are marked with “skip” in the table.

### 3.1.2 Idiomatic classes

The narrowed list of phrases, which I further analyzed, contains 131 expressions (the complete table is appended at the end of this thesis). In my sample, most frequently used colour is black, followed by blue, red, white and green. Very few phrases containing the word yellow and grey (gray) were found. The last two colours, namely violet and orange, are virtually non-existent in idioms.

I identified the five idiomatic classes, as described by Tkatchuk (see section 2.1.2). In this sample, the most frequent were nominal idioms. I found that rather surprising, because in Tkatchuk’s work, 180 out of 260 idioms were verbal idioms, nominal accounted for only 36 items. However, as Tkatchuk noted, “it has to be taken into consideration that there are many nominal, adjectival and adverbial idioms derived
from the verbal ones” (65). Colour idioms analyzed in this work were classified according to their actual form and no attempt was made to find out whether they are derived from some other class or not. Below is a complete table listing the numbers of occurrences of the five classes for each colour (4 phrases were included in two classes at the same time, because it was not possible to distinguish their instances - for example *blackleg* was analyzed as one expression, although it can be a noun as well as a verb).

<table>
<thead>
<tr>
<th></th>
<th>nominal</th>
<th>adjectival</th>
<th>adverbial</th>
<th>verbal</th>
<th>sentential</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>17</td>
<td>10</td>
<td>4</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>blue</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>red</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>green</td>
<td>4</td>
<td>3</td>
<td>-</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>white</td>
<td>8</td>
<td>9</td>
<td>-</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>yellow</td>
<td>4</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>grey</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>violet</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>orange</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>total</td>
<td>52</td>
<td>35</td>
<td>15</td>
<td>28</td>
<td>5</td>
</tr>
</tbody>
</table>

*Table 4: Idiomatic classes of the expressions in the sample*

### 3.1.3 Distribution of analyzed phrases in the corpora

Table 5 below provides a summary of occurrences of the chosen idioms in all three corpora used.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>8221</td>
<td>1390</td>
<td>1933</td>
<td>1951</td>
<td>2032</td>
<td>915</td>
<td>1737</td>
<td>194</td>
<td>524</td>
<td>224</td>
<td>122</td>
<td>673</td>
<td>3897</td>
</tr>
<tr>
<td>blue</td>
<td>2758</td>
<td>584</td>
<td>490</td>
<td>671</td>
<td>757</td>
<td>256</td>
<td>226</td>
<td>12</td>
<td>35</td>
<td>45</td>
<td>11</td>
<td>123</td>
<td>859</td>
</tr>
<tr>
<td>red</td>
<td>2001</td>
<td>444</td>
<td>387</td>
<td>467</td>
<td>542</td>
<td>161</td>
<td>519</td>
<td>36</td>
<td>135</td>
<td>122</td>
<td>27</td>
<td>199</td>
<td>1411</td>
</tr>
<tr>
<td>white</td>
<td>1778</td>
<td>284</td>
<td>262</td>
<td>448</td>
<td>520</td>
<td>264</td>
<td>414</td>
<td>8</td>
<td>71</td>
<td>56</td>
<td>132</td>
<td>147</td>
<td>949</td>
</tr>
<tr>
<td>green</td>
<td>697</td>
<td>134</td>
<td>127</td>
<td>205</td>
<td>177</td>
<td>54</td>
<td>99</td>
<td>4</td>
<td>9</td>
<td>30</td>
<td>1</td>
<td>55</td>
<td>237</td>
</tr>
<tr>
<td>yellow</td>
<td>111</td>
<td>12</td>
<td>16</td>
<td>31</td>
<td>33</td>
<td>33</td>
<td>18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>97</td>
</tr>
<tr>
<td>grey</td>
<td>637</td>
<td>137</td>
<td>63</td>
<td>191</td>
<td>152</td>
<td>94</td>
<td>179</td>
<td>12</td>
<td>14</td>
<td>28</td>
<td>38</td>
<td>87</td>
<td>183</td>
</tr>
<tr>
<td>violet</td>
<td>52</td>
<td>9</td>
<td>8</td>
<td>18</td>
<td>16</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>orange</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>16255</td>
<td>2994</td>
<td>3286</td>
<td>3970</td>
<td>4227</td>
<td>1778</td>
<td>3199</td>
<td>267</td>
<td>793</td>
<td>515</td>
<td>332</td>
<td>1292</td>
<td>7647</td>
</tr>
</tbody>
</table>

26
Table 5: The numbers of occurrences of analyzed phrases in corpora sections

Each line shows totals for the given colour in the separate sections as well as in the corpus as a whole. As explained in Chapter 2, the TIME is not analyzed by sections. The last line contains totals for all the analyzed phrases in each register.

In further discussion, the BNC’s section “Miscellaneous” (“Misc” in the table) will be ignored, because it contains texts from various sources, not just one specific genre, and it is larger than the remaining sections (see 2.2.4) – this disproportion would certainly distort the results.

3.2 Black

What is quite obvious after looking at the phrases listed in the appended table is that black colour is not used to convey positive meanings. There are some neutral phrases, which do not suggest any negative evaluation or connotation, such as put something down in black and white or the adjective black-tie, but these are rather scarce and they are mostly related to the colour of described objects, i.e. the motivation of the phrase is quite clear. Most of the phrases in the group are associated with disapproval (black sheep of the family, black look), sadness and gloom (black mood), illegality (black economy, blackmail) and evil (black deeds, black as hell).

This group of idioms is the largest one in my sample. It contains 43 phrases: 17 nominal, 14 verbal, 10 adjectival, four adverbial, and one sentential; two of these phrases were categorized as nominal/verbal and one as nominal/adjectival, the reason for that was provided in one of the previous sections. In the actual analysis, the adverbial phrase in the black had to be skipped, because the corpora did not provide valid data (see section 3.1.1). Some of the other popular expressions also need to be looked at with some benevolence, because as their numbers of occurrences reached several hundred, it was not feasible to go through every single one of them and check
the contexts to see whether it is the meaning looked for or not. This applies mainly to
the expressions *black-and-white* and *in black and white*, which referred frequently
simply to the colour of the described objects (photographs, movies, etc.). The phrase
*black-tie* exists also without hyphen; however, the results yielded by corpora were not
relevant (i.e. they were not idiomatic) and, therefore, only the hyphenated adjective was
included in the analysis.

### 3.2.1 Frequencies in sections

<table>
<thead>
<tr>
<th></th>
<th>COCA</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>BNC</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>8221</td>
<td>1390</td>
<td>1933</td>
<td>1951</td>
<td>2032</td>
<td>915</td>
<td>1737</td>
<td></td>
<td>194</td>
<td>524</td>
<td>224</td>
<td>122</td>
</tr>
<tr>
<td>100%</td>
<td>16.9%</td>
<td>23.5%</td>
<td>23.7%</td>
<td>24.7%</td>
<td>11.1%</td>
<td>100%</td>
<td></td>
<td>11.2%</td>
<td>30.2%</td>
<td>12.9%</td>
<td>7.0%</td>
</tr>
</tbody>
</table>

*Table 6: Total numbers of occurrences of expressions with black and their proportional
distribution in sections*

When analyzed as a whole, the expressions containing the word *black* were most often
found in newspaper, magazine, and fiction sections of the COCA (the differences in
numbers are rather insignificant), least frequently in academic genre. Strangely enough,
when we look at the individual idiomatic phrases one by one, we see that the majority of
them is commonest in fiction (more than half of the analyzed idioms); and the same still
holds true, even if we take into account only those expressions that appear more than 40
times in the COCA, although the difference is not so substantial any more.

As for the BNC, the results are consistent with the American corpus: fiction,
with its 30.2% of occurrences, is by far the most frequent source of idiomatic phrases,
academic texts provide fewest tokens – not only when the totals of sections are
compared, but also when the individual phrases’ frequencies come into question (22
expressions are most often found in fiction, while only two are commonest in academic
genre).
Several idioms from this group are extremely popular according to the three web corpora: the verb to black out and the derived noun blackout, black-and-white, in black and white, blackmail, and black market all yielded more than 500 tokens in the COCA. Only black market is most often found in academic genre, which could be attributed to the probable existence of a large number of economic and legal articles dealing with this phenomenon. I did not distinguish any pattern that would suggest a connection of a certain field of meaning (for example, phrases connected with economy or illegality) with a certain corpus section. Only the similes (e.g. black as coal) seem to be associated particularly with one register: fiction.

3.3 Blue

Blue is the colour of sky, in connection to human being it suggests physical discomfort and sadness, melancholy. In the analyzed sample, several related expressions are indeed connected with the sky and its blue colour: out of the blue (the most frequent one), out of a clear blue sky, (like) a bolt out of the blue and a bolt from the blue. All of these have the same meaning – suddenly, without warning, unexpectedly – and their motivation is rather obvious: no one expects a thunderstorm, when there are no clouds in the sky. The only phrase denoting sadness that I found and analyzed was the expression to get/have the blues. Although my data do not show it (no other idiom from this semantic field was analyzed), blue colour is strongly associated with sadness, even depression. Dictionaries commonly list these feelings as another meaning of the word blue: for example, the Macmillan dictionary defines the polysemous adjective “blue” as 1. a colour of “the sky on a clear sunny day” and 2. an informal expression meaning “feeling rather sad” (140), and so does also the online Cambridge Dictionary, to name but a few.
There were eight adverbial, seven nominal, six adjectival, four verbal, and one sentential idiom in this group; one was labelled simultaneously as nominal and adjectival (*bluestocking*). I do not recall any issues arising during the data collection process that would be worth mentioning in relation to the phrases in this group.

### 3.3.1 Frequencies in sections

<table>
<thead>
<tr>
<th></th>
<th>COCA</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2758</td>
<td>584</td>
<td>490</td>
</tr>
<tr>
<td>100%</td>
<td>21.2%</td>
<td>17.8%</td>
</tr>
</tbody>
</table>

*Table 7: Total numbers of occurrences of expressions with blue and their proportional distribution in sections*

The distribution of phrases in different registers is rather balanced in the COCA (newspapers being slightly more popular than the rest of the sources); only the academic section shows significantly lower numbers (only 9.3 % of tokens). Similarly to the group of “black idioms”, although it is not the largest section on the whole, the majority of phrases is represented most often in fiction. However, if the less frequent expressions (fewer than 40 instances) are omitted, the differences are effaced and the newspaper, magazine, and fiction sections are equal again.

The British corpus contained significantly fewer occurrences of idiomatic phrases: while it consists of only four times fewer words than the COCA, it yielded 12 times less tokens. In other words, the idioms were three times more popular in the COCA than they were in the BNC, regardless the size of the respective tool. The BNC’s most and least popular sections are the same as in the COCA, both when the total numbers and the individual expressions are concerned.

Only two phrases from this group of idioms were represented more than 500 times in the COCA: the adverbial phrase *out of the blue* (778 tokens) and the adjective
blue-collar (1320 tokens); the former is most common in fiction, the latter in newspapers. There does not seem to be any pattern or connection between the denoted things and the registers, where the phrases are used most often.

3.4 Green

As far as the meaning of the phrases which contain the word green is concerned, there are two expressions that clearly support Steinvall's claim about green's association with anger in general, jealousy and envy specifically. These are green with envy and the green-eyed monster. One expression – green around/about/at the gills – fits Vaňková’s claim that in connection with human beings, green means illness or death. The remaining expressions in the group are not related to feelings or emotions and have positive connotations: e.g. the rub of the green (i.e. good luck) and give/get the green light. Several idioms refer to the green colour of plants (greener pastures, the grass is always greener on the other side, green as grass).

There were 12 phrases in my sample that belong to this group: four nominal, four verbal, three adjectival and one sentential. Most of the phrases were rather infrequent, with only a few tokens - especially the two smaller corpora provided very low numbers.

3.4.1 Frequencies in sections

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>697</td>
<td>134</td>
<td>127</td>
<td>205</td>
<td>177</td>
<td>54</td>
<td>99</td>
<td>4</td>
<td>9</td>
<td>30</td>
<td>1</td>
<td>55</td>
<td>237</td>
</tr>
<tr>
<td>100%</td>
<td>19,2%</td>
<td>18,2%</td>
<td>29,4%</td>
<td>25,4%</td>
<td>7,7%</td>
<td>100%</td>
<td>4,0%</td>
<td>9,1%</td>
<td>30,3%</td>
<td>1,0%</td>
<td>55,6%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 8: Total numbers of occurrences of expressions with green and their proportional distribution in sections
As can be easily seen in the table above, magazine and newspaper articles are frequently enlivened with idiomatic expressions. In the COCA, more than half of all the instances of idioms were found in the magazine and newspaper sections; the BNC’s newspaper section contained 30.3% of the occurrences. Idioms from this group are less often in speech and in fiction, and occur only sparingly in academic texts.

The individual expressions are again most frequently found in the register of fiction in both corpora, although in the BNC, it shares the first place with newspapers. None of the expressions is most common in academic texts. Only four phrases were represented more than 40 times in the COCA: green with envy, greener pastures and greenhorn were most frequent in magazines, give someone/something the green light in spoken language. The lastly mentioned expression was also the most frequent one in this group (276 tokens).

Two phrases from the verbal group were geographical variants of the expression meaning “to be good at growing plants”: have green fingers and have a green thumb. Americans are said to prefer the version with “thumb”, while British speakers use “fingers” instead. This was also confirmed by the results obtained from the corpora, although the numbers were very low: the COCA contained only 1 instance of have green fingers, but 17 instances of have a green thumb. Contrariwise, the BNC did not include any occurrence with “thumb”, but four with “fingers”. The TIME contained more tokens of the American version of the phrase than the British one.

There was only one “metaphorical” expression in this group: the simile (as) green as grass, which was found twice in fiction and once in academic texts in the COCA; the TIME corpus – that is comparable to the magazine or newspaper sections because of its sources – contained three tokens, the BNC one (in the ignored miscellaneous section).
3.5. Grey/Gray

Except for the phrase *gray matter*, where the colour may be attributed simply to the colour of neurons and brain in general, the expressions in this group suggest that the situation or entity that they connote is somewhat shady, illegal or immoral. *Grey area* is a situation, where the rules are unclear and it is difficult or impossible to say what is right and what is wrong. *Grey eminence* is a figure, which stays in the background and is almost invisible, but which has immense power and influence when decisions are made. Also, such persons may be associated with intrigues and plots. The last of these expressions is *grey market*; it is defined as a business of buying and selling, which is not illegal, but is considered morally wrong.

With this colour, both spellings were checked. Although “gray” is used in the American variety and “grey” in the British variety of English, both were found in the COCA and the TIME. Nevertheless, the frequency of the British “grey” in the two American corpora was rather low. On the other hand, the BNC seems to strictly obey the distinction and includes only the British spelling. Although I found some suggestions on the web that Americans might tend to use “gray” for the colour and “grey” as the adjective, my findings do not support such claim: the COCA was very consistent in that it contained more occurrences of the phrases with the spelling “gray” than with “grey” in every single case. This is most visible in the frequencies of expressions *grey/gray area* and *grey/gray matter* (28 to 410 and 7 to 123 respectively). The TIME was not as consistent: the two above-mentioned phrases are still represented in a similar way, although the ratio is significantly lower (7 to 42 and 18 to 33); but other expressions show something completely different. For example the phrase *grey/gray market* is almost ten times more frequent with the “e” variant.
If we leave the actual numbers aside and look only at the distribution of the phrases in the TIME over years, we will find an interesting pattern – the expressions with the British spelling are to be found almost exclusively in the sections from the 1920s to the 1960s, while the spelling with “a” occurs from the 1970s onwards. According to the article “How do you spell the color grey or gray?”, “grey” became the established British spelling in the 20th century. However, in the United states, “gray” seems to have become standard earlier already: both Webster's Academic Dictionary (1867) and Webster's Condensed Dictionary (~1897) “include entries for “grey” that refer readers to “gray” for the term's definitions”. Nevertheless, judging from the numbers provided by the TIME corpus, the geographical differences in usage have not settled until around 1960s/1970s, when the Americans (or at least the TIME magazine contributors) virtually stopped using the version with “a”.

<table>
<thead>
<tr>
<th>grey area(s)</th>
<th>1920s</th>
<th>1930s</th>
<th>1940s</th>
<th>1950s</th>
<th>1960s</th>
<th>1970s</th>
<th>1980s</th>
<th>1990s</th>
<th>2000s</th>
</tr>
</thead>
<tbody>
<tr>
<td>gray area(s)</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grey eminence(s)</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gray eminence(s)</td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grey market(s)</td>
<td></td>
<td></td>
<td>24</td>
<td>21</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gray market(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>grey matter(s)</td>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gray matter(s)</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 9: Distribution of grey/gray over years in the TIME corpus

3.5.1 Frequencies in sections

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>637</td>
<td>137</td>
<td>63</td>
<td>191</td>
<td>152</td>
<td>94</td>
<td>179</td>
<td>12</td>
<td>14</td>
<td>28</td>
<td>38</td>
<td>87</td>
<td>183</td>
</tr>
<tr>
<td>100%</td>
<td>21.5%</td>
<td>9.9%</td>
<td>30.0%</td>
<td>23.9%</td>
<td>14.8%</td>
<td>100%</td>
<td>6.7%</td>
<td>7.8%</td>
<td>15.6%</td>
<td>21.2%</td>
<td>48.6%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 10: Total numbers of occurrences of expressions with grey/gray and their proportional distribution in sections
All of the four analyzed expressions belong to the nominal class. On the whole, most tokens in the COCA were in the magazines and newspaper sections. *Gray area* is the most popular idiom in this group and it is most frequently found in spoken English, followed very closely by newspapers and then magazines; only about 15% of the occurrences is from the academic section, 5% from the fiction section. *Gray eminence* is not a very frequent expression, but more than one third of it is to be found in fiction. Nevertheless, the TIME corpus, which is comparable to the magazines and newspaper sections of the COCA, contains 25 occurrences of this expression (both spellings included). The last two phrases – *gray market* and *gray matter* – are most popular in magazines in the COCA; their British counterparts in newspapers and fiction respectively in the BNC.

### 3.6 Orange

I found several expressions with orange colour; however, only one proved to be a pure idiom: *(all) Lombard street to China orange*. The remaining expressions collected during the preliminary search in dictionaries were excluded from the further corpora search, as explained in Chapter 2. The idiom *(all) Lombard street to China orange* is described as old-fashioned, but no information was found as to when it was coined. It is not listed in the table of idioms appended to this work, because none of the Davies’ corpora contained this expression, which probably means that the phrase is either very old or not spread and used enough. The admittance of this phrase into the group of colour idioms would be questionable anyway, as it refers to the citrus fruit, not the colour as such.
3.7 Red

Although people certainly associate love, affection and passion with the red (and pink) colour, red is by far most frequently connected with anger, according to Steinvall’s findings. In the sample gathered for the purpose of this paper, only see red and red rag to a bull fit the anger category. Red also serves to describe embarrassment, like in the verbal expression go red. There were also several adjectival phrases with similar meaning, which were based on the person’s facial colour. These adjectives were listed as separate expressions in the table, although they are very closely related: beetroot red is derived from red as a beetroot, beet red from red as a beet. There does not seem to be any connection between the other expressions that would make it possible to categorize them and ascribe the red colour a specific meaning. The nominal phrase red ink and adverbials in/out of the red refer, obviously, to the colour used by financial institutions on statements and reports. In a similar way, red carpet treatment is motivated by the red colour of the carpet that is traditionally rolled out for VIP guests and famous or important persons at certain events and places, such as presidential visits or popular film festivals.

From the 21 expressions admitted for analysis in this group, seven were nominal, five adjectival, five verbal, three adverbial, and one sentential. It may be argued, that the phrase red in tooth and claw is not sentential, but adjectival. I categorized it as sentential, because it is usually used in apposition to the word “nature” (i.e. nature, red in tooth and claw), which, then, has the sound of a proverbial phrase.

An interesting pattern was discerned in the distribution of the expression red ink. While it referred to economy – to the unpleasant situation of being in debt – in the newspaper section, in fiction genre the instances of red ink were often simple descriptions of the colour (e.g. “Whenever he came across one, he underlined it
in red ink and dog-eared the page so he could find it again in a hurry.” or “I mumbled, hands stained with red ink.” – COCA). The numbers of tokens of the verbal idiom go red may not be absolutely precise, because sometimes, this phrase was not referring to the colour of someone's face turning red because of embarrassment, but to the changing colour of the leaves in autumn or the colour of the sky during the sunset.

### 3.7.1 Frequencies in sections

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>444</td>
<td>387</td>
<td>467</td>
<td>542</td>
<td>161</td>
<td>519</td>
<td>36</td>
<td>135</td>
<td>122</td>
<td>27</td>
</tr>
<tr>
<td>100%</td>
<td>22,2%</td>
<td>19,3%</td>
<td>23,3%</td>
<td>27,1%</td>
<td>8,0%</td>
<td>100%</td>
<td>6,9%</td>
<td>26,0%</td>
<td>23,5%</td>
<td>5,2%</td>
</tr>
</tbody>
</table>

*Table 11: Total numbers of occurrences of expressions with red and their proportional distribution in sections*

Generally speaking, the newspaper section is the most frequent source of occurrences of “red idioms” in the COCA and second most frequent in the BNC (first being the genre of fiction). However, when we look at the frequencies of individual phrases, we see that the results are different: seven phrases yield highest numbers of occurrences in fiction, six in magazines, five in spoken language, three in newspapers and one in the academic section. If the less frequent expressions are excluded, of the nine phrases that are represented more than 40 times in the COCA, only the expression red tape and red ink were most often found in newspapers; these two phrases were also the most frequent “red idioms” in the corpus (with 667 and 415 tokens respectively). The phrases most popular in magazines are out of the red, red alert and catch someone red-handed; the last one being similarly frequent in fiction as well. The fiction genre contains most occurrences of the descriptive idioms beet red and the verb to go red. The other phrases denoting embarrassment are not used very often; except for the already mentioned
phrase *beet red*, the adjectives were not found more than five times in any of the corpora used.

It was interesting to see the British–American English distinction at work again, although, in this case, the results might be assigned to a mere coincidence, as the numbers are extremely low: according to the Cambridge dictionary online, the adjective *beet red* is American, while *beetroot red* is British. Keeping with that, the BNC contained only the British version of the phrase, the TIME and the COCA only the American version. No other phrase with geographical variation was found.

This phrasal group contained two similes: *as red as a beetroot* and *as red as beet*. They were represented only in the American corpus, the former was found twice in fiction, the latter four times in fiction and once in magazines.

### 3.8 Violet

For this colour, only one idiomatic expression seems to exist in English: the nominal phrase *shrinking violet*. According to Answers.com, this expression was coined sometime in the early 1900s. In the TIME, it appears in the 1930s for the first time.

#### 3.8.1 Frequencies in sections

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>9</td>
<td>8</td>
<td>18</td>
<td>16</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>100%</td>
<td>17,3%</td>
<td>15,4%</td>
<td>34,6%</td>
<td>30,8%</td>
<td>1,9%</td>
<td>100%</td>
<td>14,3%</td>
<td>14,3%</td>
<td>42,9%</td>
<td>0,0%</td>
<td>28,6%</td>
<td>100%</td>
</tr>
</tbody>
</table>

*Table 12: Total numbers of occurrences of “shrinking violet” and their proportional distribution in sections*

The phrase is classified as informal in various sources, which is supported also by the distribution of occurrences in the COCA, the BNC, and the TIME. Even though the total numbers are rather low (52, 7, and 14 respectively), the phrase is much more
common in newspapers and magazines, while only one occurrence in all three corpora falls into the academic category.

3.9 White

As was already mentioned in the section devoted to colour semiotics, according to Flavell, in English, white indicates good and it has generally positive connotations (30). That is an interesting note, but it cannot be taken as a rule that would help language learners understand idiomatic phrases containing the word white. My sample included quite a large number of expressions (considering the size of the material) which were rather negative. There is, undoubtedly, hardly anything positive about being white-livered, bleeding someone white, or behaving like a whited sepulchre. And if you had a white elephant or experienced several white nights in a row, you would not find it pleasant either. Moreover, quite in opposition to Flavell’s claim, Steinvall associates white with anger and fear, when emotions and feelings come into play (e.g. white-knuckle(d) or the above mentioned white-livered).

In this group of idioms, eight were nominal, nine adjectival, one verbal and one sentential. The phrase white night was represented quite frequently in the COCA, but the analysis revealed that there were many occurrences from Rajnar Vajra's book Viewschool, where White Night was the name of a character. This character accounted for 31 occurrences of the phrase in the fiction section of the corpus, which were not included in the table. At first sight, the expression the white man’s burden, which is defined by Cowie as “the former concept of the European's duty to advance civilization, education, trade, public health, etc in underdeveloped parts of the world”, seems to be American, because the BNC contains only four instances. However, a lot of the occurrences refer only to Rudyard Kipling's poem of the same title. In this case, the irrelevant instances of the phrase were not filtered out and remained in the table,
because it was not always absolutely clear from the context, whether the phrase refers to
the poem’s title or to the concept.

3.9.1 Frequencies in sections

<table>
<thead>
<tr>
<th></th>
<th>COCA</th>
<th>BNC</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1778</td>
<td>284</td>
<td>262</td>
<td>448</td>
</tr>
<tr>
<td>100%</td>
<td>16,0%</td>
<td>14,7%</td>
<td>25,2%</td>
</tr>
</tbody>
</table>

Table 13: Total numbers of occurrences of expressions with white and their proportional distribution in sections

As can be seen in Table 13 above, the COCA contained most tokens of the analyzed phrases in the newspaper and magazine sections; the remaining three sections were less frequent sources of idiomatic expressions and were represented by almost equal numbers of tokens. The TIME provided surprisingly high number of idiomatic phrases, considering the size of the corpus: the expressions in the smaller American corpus are twice as popular as in the COCA. The BNC yielded very different results from those of the COCA: the academic section contained almost one third of all the tokens in the corpora. However, these numbers are greatly affected by the popularity of the adjective white-collar, which accounts for more than half of the occurrences in all three corpora (996 in the COCA, 228 in the BNC, and 511 in the TIME). In the BNC, 121 instances of this phrase are found in the academic section. No other expression containing the word “red” was similarly popular. From the eleven expressions that were found in the COCA more than 40 times, five were most frequent in fiction (two of them were the similes as white as a sheet and as white as snow), four in newspapers, one in magazines and one in the academic section (the white man's burden).
3.10 Yellow

Yellow colour was not represented in idiomatic phrases very often. From the expressions collected in the first part of the work, only a few were accepted as idioms. In fact, there were only four separate idioms, all of them belonging to the nominal class: yellow journalism, the semantically related yellow press, Yellow Peril and yellow-belly.

The last phrase had also derived adjectival versions, one with hyphen and one without hyphen.

Connection of yellow with a certain meaning in my sample is hard to prove. Steinvall and Vaňková link yellow colour with joy, but also with jealousy, envy, and hatred. Besides these emotional states, it is also associated with old age and long-term illness. Neither of them connects yellow with fear, which is the only emotion traceable in the expressions analyzed in this work (yellow-belly). The expression Yellow Peril (fear that the Chinese and/or Japanese nations will overrun the world) is apparently motivated by the colour of complexion of Chinese and Japanese people. The phrase yellow journalism is a term used for colourful, sensational reporting, supplemented in newspapers with “banner headlines, coloured comics, and copious illustrations”. The expression “was coined in the 1890s to describe the tactics employed in furious competition between two New York City newspapers, the World and the Journal”. The source for the term was the immensely popular cartoon The Yellow Kid drawn by Richard F. Outcault (for the World, later the Journal) and George B. Luks (the World) (“Yellow journalism”).
3.10.1 Frequencies in sections

<table>
<thead>
<tr>
<th></th>
<th>COCA 111</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>BNC 18</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>TIME 97</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>16</td>
<td>19</td>
<td>31</td>
<td>33</td>
<td></td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>10,8%</td>
<td>14,4%</td>
<td>17,1%</td>
<td>27,9%</td>
<td>29,7%</td>
<td>100%</td>
<td>0,0%</td>
<td>22,2%</td>
<td>38,9%</td>
<td>5,6%</td>
<td>33,3%</td>
</tr>
</tbody>
</table>

*Table 14: Total numbers of occurrences of expressions with yellow and their proportional distribution in sections*

The noun *yellow-belly* and the adjective *yellow-bellied* (or *yellowbellied*) were found only a few times in the corpora. Of these three related expressions, the BNC contained only the hyphenated adjective. The term without hyphen was found only in the COCA, which might be attributed to the tendency of American English to simplify spellings; however, the number of occurrences is too low to provide sufficient evidence and support such conclusions. Moreover, the TIME did not contain this expression (without hyphen) at all.

In general, the BNC included significantly fewer phrases with the word “yellow” than the other two corpora. Even though the TIME and the BNC are similar as to their sizes, the TIME contained 5 times as many occurrences of the listed idiomatic expressions, which might lead to the assumption that American English is more idiomatic than British English. The numbers of occurrences in the individual sections can be seen in table 14 above. In the BNC, the newspaper section was the most frequent source of idioms containing the word yellow. The COCA contained most items in the academic section, followed closely by the newspaper section. However, the individual idioms were not most popular in one specific section - *yellow press* and *yellow journalism* appeared most frequently in newspapers, the two versions of *yellow-bellied* (with and without hyphen) in fiction, *Yellow Peril* in academic texts. The numbers of
tokens were too low, with only one expression reaching more than 40, and no specific pattern was recognized.

3.11 Similes in English

Similes are well-known figures of speech used frequently in literature because of their expressiveness. Wikberg defines simile as “a figurative expression used to make an explicit comparison of two unlike things by means of the prepositions like, (as)…as or the conjunctions as, as if, as though” (156).

My sample of idiomatic phrases included similes as well, even though they are considered to be borderline cases between idioms and collocations. There were eleven of them: black as coal, black as ink, black as hell, black as the devil, black as pitch, red as a beetroot, red as a beet, white as sheet, white as snow, white as chalk, and green as grass. Except for the expressions white as sheet (48 tokens) and white as snow (43 tokens) they were found only a few times, rarely reaching more than 10 instances in the largest of the used corpora, the COCA. All of these phrases were found most often in the fiction sections of both the American and the British corpus. The percentages were very high and some expressions were not present in any other section of the corpora at all. These findings are consistent with Wigberg’s statement that “similes are figurative and it is no surprise that they are presented as most frequent in literary discourse” (140).

3.12 Idiomatic expressions over time

In this subchapter, I would like to at least briefly touch upon the idea that idiomatic expressions change over time. I will focus only at the changes in popularity, if any; changes in meaning will not be investigated here. Initially, I had been planning to base my work on the differences in the popularity of idioms in different time periods. However, I decided against it, because even those phrases, which I considered well-
known and popular, were providing too low numbers to enable me to compare them in time. Here, I will explore only the expressions represented more than 400 times in the COCA; I believe that such numbers should provide sufficient data for the analysis.

There were only 12 phrases in the whole sample that passed my frequency quota: they are listed in the Table 15 below. The numbers in the first line stand for years: the COCA is divided into sections by five years (“90” indicates the period from 1990 to 1994, etc.), the TIME is divided into decades (“20s” stands for the 1920s, etc.). The numbers of occurrences in the COCA may not be exactly the same as those listed in the appended table of idioms, because there are some discrepancies in the corpus: it seems as if some of the tokens were not tagged and categorized properly and therefore they do not appear in the results for some genres or periods. Sometimes, the total of genre sections is lower than that of time periods; sometimes it is the other way round. No explanation for this issue was found, though.

<table>
<thead>
<tr>
<th>phrase</th>
<th>COCA</th>
<th>90</th>
<th>95</th>
<th>00</th>
<th>05</th>
<th>20s</th>
<th>30s</th>
<th>40s</th>
<th>50s</th>
<th>60s</th>
<th>70s</th>
<th>80s</th>
<th>90s</th>
<th>00s</th>
</tr>
</thead>
<tbody>
<tr>
<td>black out</td>
<td>567</td>
<td>118</td>
<td>151</td>
<td>122</td>
<td>176</td>
<td>252</td>
<td>-</td>
<td>11</td>
<td>67</td>
<td>55</td>
<td>45</td>
<td>27</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>blackout(s)</td>
<td>1246</td>
<td>176</td>
<td>199</td>
<td>603</td>
<td>268</td>
<td>820</td>
<td>-</td>
<td>43</td>
<td>277</td>
<td>63</td>
<td>133</td>
<td>148</td>
<td>66</td>
<td>27</td>
</tr>
<tr>
<td>black-and-white</td>
<td>2428</td>
<td>603</td>
<td>590</td>
<td>719</td>
<td>516</td>
<td>472</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>63</td>
<td>98</td>
<td>53</td>
<td>90</td>
<td>95</td>
</tr>
<tr>
<td>in black and white</td>
<td>624</td>
<td>155</td>
<td>180</td>
<td>152</td>
<td>137</td>
<td>122</td>
<td>16</td>
<td>8</td>
<td>3</td>
<td>14</td>
<td>31</td>
<td>18</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>blackmail</td>
<td>987</td>
<td>316</td>
<td>252</td>
<td>243</td>
<td>176</td>
<td>758</td>
<td>28</td>
<td>77</td>
<td>77</td>
<td>136</td>
<td>131</td>
<td>150</td>
<td>94</td>
<td>50</td>
</tr>
<tr>
<td>black market(s)</td>
<td>986</td>
<td>426</td>
<td>203</td>
<td>189</td>
<td>168</td>
<td>625</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>269</td>
<td>111</td>
<td>76</td>
<td>62</td>
<td>46</td>
</tr>
<tr>
<td>out of the blue</td>
<td>782</td>
<td>172</td>
<td>209</td>
<td>232</td>
<td>169</td>
<td>67</td>
<td>-</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>14</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>blue-collar</td>
<td>1334</td>
<td>403</td>
<td>311</td>
<td>320</td>
<td>300</td>
<td>475</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>60</td>
<td>170</td>
<td>123</td>
<td>73</td>
</tr>
<tr>
<td>red ink</td>
<td>423</td>
<td>142</td>
<td>86</td>
<td>120</td>
<td>75</td>
<td>358</td>
<td>8</td>
<td>21</td>
<td>35</td>
<td>36</td>
<td>51</td>
<td>54</td>
<td>92</td>
<td>49</td>
</tr>
<tr>
<td>red tape</td>
<td>680</td>
<td>219</td>
<td>169</td>
<td>145</td>
<td>147</td>
<td>579</td>
<td>21</td>
<td>37</td>
<td>197</td>
<td>90</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>26</td>
</tr>
<tr>
<td>white-collar</td>
<td>1049</td>
<td>308</td>
<td>269</td>
<td>257</td>
<td>215</td>
<td>511</td>
<td>2</td>
<td>43</td>
<td>56</td>
<td>77</td>
<td>88</td>
<td>84</td>
<td>78</td>
<td>42</td>
</tr>
<tr>
<td>gray area(s)</td>
<td>411</td>
<td>95</td>
<td>94</td>
<td>124</td>
<td>98</td>
<td>42</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>7</td>
<td>15</td>
<td>6</td>
</tr>
</tbody>
</table>

*Table 15: The most frequent idiomatic phrases and their numbers of occurrences over time*
The distribution of words in the corpora in time is not absolutely even (to compare with division into registers see section 2.2); the COCA is represented in the following way:

- 1990-1994: 103.3 million
- 1995-1999: 102.9 million
- 2000-2004: 102.6 million
- 2005-2005: 93.6 million

The last section that contains the most recent texts is about 10% smaller than the rest, which needs to be kept in mind when comparing the actual numbers. The distribution in the TIME is much more problematic, because the sections’ sizes are not comparable at all, as some of the smaller sections contain less than half the amount of words of the largest ones. The actual numbers are:

- 1920s: 7.6 million
- 1930s: 12.7 million
- 1940s: 15.4 million
- 1950s: 16.8 million
- 1960s: 16.1 million
- 1970s: 13.6 million
- 1980s: 11.4 million
- 1990s: 9.7 million
- 2000s: 6.4 million

The only appropriate way of assessing the popularity of the phrases in the TIME corpus seems to be the use of the “per million” figure, which takes into account the different sizes of the individual sections. The table below contains the most popular expressions
with their numbers of occurrences transformed into the “per million” figures, which can be accessed in the corpus interface.

<table>
<thead>
<tr>
<th></th>
<th>COCA 90</th>
<th>COCA 95</th>
<th>COCA 00</th>
<th>COCA 05</th>
<th>TIME 20s</th>
<th>TIME 30s</th>
<th>TIME 40s</th>
<th>TIME 50s</th>
<th>TIME 60s</th>
<th>TIME 70s</th>
<th>TIME 80s</th>
<th>TIME 90s</th>
<th>TIME 00s</th>
</tr>
</thead>
<tbody>
<tr>
<td>black out</td>
<td>1.14</td>
<td>1.47</td>
<td>1.19</td>
<td>1.88</td>
<td>-</td>
<td>0.87</td>
<td>4.34</td>
<td>3.28</td>
<td>2.80</td>
<td>1.99</td>
<td>2.02</td>
<td>1.54</td>
<td>1.40</td>
</tr>
<tr>
<td>blackout(s)</td>
<td>1.70</td>
<td>1.93</td>
<td>5.88</td>
<td>2.86</td>
<td>-</td>
<td>3.40</td>
<td>17.92</td>
<td>3.75</td>
<td>8.27</td>
<td>10.89</td>
<td>5.80</td>
<td>2.77</td>
<td>9.80</td>
</tr>
<tr>
<td>black-and-white</td>
<td>5.84</td>
<td>5.73</td>
<td>7.01</td>
<td>5.52</td>
<td>1.05</td>
<td>0.16</td>
<td>0.26</td>
<td>3.75</td>
<td>6.09</td>
<td>3.90</td>
<td>7.91</td>
<td>9.76</td>
<td>9.18</td>
</tr>
<tr>
<td>in black and white</td>
<td>1.50</td>
<td>1.75</td>
<td>1.48</td>
<td>1.46</td>
<td>3.67</td>
<td>0.63</td>
<td>0.19</td>
<td>0.83</td>
<td>1.93</td>
<td>1.32</td>
<td>0.88</td>
<td>1.34</td>
<td>1.40</td>
</tr>
<tr>
<td>blackmail</td>
<td>3.06</td>
<td>2.45</td>
<td>2.37</td>
<td>1.88</td>
<td>3.67</td>
<td>6.08</td>
<td>4.98</td>
<td>8.10</td>
<td>8.15</td>
<td>11.04</td>
<td>8.27</td>
<td>5.14</td>
<td>2.33</td>
</tr>
<tr>
<td>black market(s)</td>
<td>4.12</td>
<td>1.97</td>
<td>1.84</td>
<td>1.80</td>
<td>-</td>
<td>-</td>
<td>17.41</td>
<td>6.61</td>
<td>4.73</td>
<td>4.56</td>
<td>4.05</td>
<td>4.01</td>
<td>3.42</td>
</tr>
<tr>
<td>out of the blue</td>
<td>1.66</td>
<td>2.03</td>
<td>2.26</td>
<td>1.81</td>
<td>-</td>
<td>0.08</td>
<td>0.58</td>
<td>0.54</td>
<td>0.56</td>
<td>0.37</td>
<td>1.23</td>
<td>1.03</td>
<td>1.56</td>
</tr>
<tr>
<td>blue-collar</td>
<td>3.90</td>
<td>3.02</td>
<td>3.12</td>
<td>3.21</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.18</td>
<td>3.73</td>
<td>12.51</td>
<td>10.82</td>
<td>7.50</td>
<td>7.16</td>
</tr>
<tr>
<td>red ink</td>
<td>1.37</td>
<td>0.84</td>
<td>1.17</td>
<td>0.80</td>
<td>1.05</td>
<td>1.66</td>
<td>2.26</td>
<td>2.14</td>
<td>3.17</td>
<td>3.97</td>
<td>8.09</td>
<td>5.03</td>
<td>1.87</td>
</tr>
<tr>
<td>red tape</td>
<td>2.12</td>
<td>1.64</td>
<td>1.41</td>
<td>1.57</td>
<td>2.75</td>
<td>2.92</td>
<td>12.75</td>
<td>5.36</td>
<td>4.97</td>
<td>5.89</td>
<td>2.64</td>
<td>2.67</td>
<td>2.80</td>
</tr>
<tr>
<td>white-collar</td>
<td>2.98</td>
<td>2.32</td>
<td>2.51</td>
<td>2.30</td>
<td>0.26</td>
<td>3.40</td>
<td>3.62</td>
<td>4.59</td>
<td>5.47</td>
<td>6.18</td>
<td>6.86</td>
<td>4.31</td>
<td>6.36</td>
</tr>
<tr>
<td>gray area(s)</td>
<td>0.92</td>
<td>0.91</td>
<td>1.21</td>
<td>1.05</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.12</td>
<td>0.52</td>
<td>1.32</td>
<td>0.62</td>
<td>1.87</td>
<td></td>
</tr>
</tbody>
</table>

Table 16: The most frequent idiomatic phrases and their numbers of occurrences over time expressed in tokens per million

When we look at the numbers in the table above closely, we can see that the popularity of the phrases changes usually gradually rather than by fits and starts. For example, the numbers (in the TIME corpus) associated with the expression red ink are increasing slowly but surely: the popularity of the phrase grows steadily from the 1920s to the 1980s (with only a slight recession in the 1950s). In the 1980s it suddenly surged, jumping from 3.97 to 8.09 tokens per million of words in the corpus, and then almost as suddenly receded – to 5.03 in the 90s, and to 1.87 in the 2000s. In the COCA, such development is not visible, because the corpus is made of texts from only the last twenty years.

Similar behaviour is recognizable in other idioms as well. The adjective blue-collar had not appeared in the TIME corpus till the 1950s; then it was gaining on popularity steadily, reaching its peak in the 70s (12.51 tokens/million) and decreasing
again gradually from 1980 onwards. The phrase is motivated by the blue colour of clothes that has been popular among the manual workers: light or navy blue work shirts and coveralls. The lower numbers of the idiom in the last two or three decades might be analogous to the decreasing number of blue-collar workers, which results from the move of production east to countries with cheaper labour.

The idiomatic noun *black market(s)* is also interesting. It appears in the TIME in the 1940s, but although it is a new expression, not found in the corpus previously at all, it is an extremely popular one all of a sudden. Ayto\(^5\) (qtd in Phrase Finder) claims that: “[a]fter a slow start in the 1930s, mainly in the area of currency dealing, the term really took off in the disrupted economic circumstances of World War II”. That would explain the extraordinary rate of 17.41 tokens/million in the first decade after the phrase's appearance. After the initial boom in the usage of the expression, it lost in popularity considerably in the 1950s (6.61 tokens/million); since then, the numbers are slightly lower each decade, which I believe to be rather surprising. It is understandable that the expression was used abundantly in the war and post-war years, when illegal selling and buying of hard-to-get and rationed goods was the issue. However, I would expect it to be used more frequently in the last decades again: although prohibition is not common any more, prostitution, drugs, and copyrighted media, to name but a few, are much more widespread.

This assumption of mine could be supported by the numbers in the COCA. The expression yielded 4.12 tokens/million in 1990-1994, a period after the end of the Cold War and the fall of the Iron Curtain. With the euphoria that emerged following that, and the easier transfer of goods across the borders, the black market could flourish (for

several years) – anything brought from the West was believed to be good and people
were willing to buy such things. This, of course, did not happen in the English speaking
countries; but the British and American media, as well as media in other countries, must
have written a great deal about the situation in Europe, as it was a major event with an
immense impact on everyone.

Could also the popularity of the noun *blackout(s)* be motivated by the
happenings in the world? One of its meanings is “a period during a war when the lights
are turned off so that an enemy cannot see them at night” – this is quite a probable
reason for the extremely high number of occurrences of the phrase in the 1940s (17.92
tokens per million). But what is behind the second peak in popularity in the 1970s
(10.89 tokens/million) and the third one in the 2000s (9.80)? Besides other meanings,
*blackout* refers also to the loss of communication that occurs when a spacecraft re-
enters the Earth's atmosphere. I assume that this meaning might be partially responsible
for the greater number of tokens of the phrase in the sixties and seventies, as the well-
known space program Apollo, followed by Skylab, was under way at that time. It would
be understandable if the media covered the communication blackouts in their reports of
the space program, because it was followed by wide public on radio and TV, especially
the extremely problematic mission of Apollo 13 in 1970. Another meaning of the phrase
*blackout* is “a situation in which journalists are officially prevented from reporting news
about something”. The most commonly known version of a media blackout is the
convention that the name of a victim of a rape or sexual assault is not made public
without the person’s consent. In the 2000s, several blackouts occurred: one was used
during 2005 New York City transit strike to enable better negotiations (“Media
blackout”); another two when a Canadian journalist (in 2008) and a New York Times
reporter (in 2009) were abducted by terrorist organizations, in order to assure the
victims’ safe return (“Reporter's escape”). The related verb to black out was (similarly to the noun) most frequent in the forties, but no other surge in popularity occurred in this case.

Not all of the listed phrases are analyzable in a similar way, though. For example, I do not see any reason, why the expression black-and-white should be most popular in the nineties, or why the verb blackmail is most frequently found in the TIME in the 1970s and the least in recent years. Moreover, the explanations of the changes in popularity are nothing more than mere speculations, as the association of a phrase with a certain situation or event is usually rather very hard to prove.

3.13 Summary

3.13.1 Popularity of idioms in American and British English

According to the results of the analysis, idioms are more popular in American English than in British English. The table in section 3.1.3 nicely shows the different figures for all three corpora. The BNC and the TIME consist of the same volume of texts (100 million words), but the TIME corpus yields 2.4 times more occurrences of idiomatic expressions than the British corpus. However, this comparison does not stand on very firm ground, because the makeup of the two corpora is not the same: while the BNC consists of texts from various sources, the TIME does not include any other genres than magazines. Nevertheless, if we compare the BNC with the third and largest corpus, the COCA, we come to a similar conclusion: the COCA is four times as large as far as the size of the material is concerned (it is composed of 400 million words), but provides 5.1 times more instances of the analyzed idioms (that is more than 16,200 to 3,200).
3.13.2 Popularity of idioms in media versus academic texts

As for the American corpus, my assumption about bigger popularity of idioms in newspapers and magazines than in academic texts was broadly valid: in total numbers, the newspaper section was indeed the largest one (26 % of the occurrences), followed by magazines (24.4 %); academic texts, on the other hand, provided the fewest instances of idiomatic expressions (10.9 %). Fiction was responsible for 20.2 % of all the instances of idiomatic phrases in the COCA.

The TIME corpus cannot be compared in a similar way, because it does not contain texts from any other sources than the TIME magazine. Still, more than 7,600 instances of phrases in such a small collection of texts may suggest the popularity of idiomatic expressions in printed media.
4 CONCLUSION

In this thesis, a specific group of idiomatic phrases chosen and collected according to certain criteria described in Chapter 2 was analyzed using three online text corpora: the Corpus of Contemporary American English, the TIME Magazine Corpus of American English, and the British National Corpus. My hypothesis was that idiomatic expressions are more frequent in newspapers and magazines than in other types of texts, academic texts in particular.

The hypothesis concerning the supposed bigger popularity of idioms in newspapers and magazines was confirmed only partially: while it is true that idiomatic expressions are more common in printed media and the least common in academic texts, the results for the fiction section were comparable. That section contained high numbers of tokens and many of the less popular phrases (with only a few tokens) were found primarily there.

Based on the numbers of tokens of the analyzed phrases in the three corpora, it seems that the American variety of English is more idiomatic than the British. Nevertheless, the results do not provide sufficient evidence to allow me to make such claims. A much larger corpus would be necessary in order to make the investigation of the popularity of idiomatic phrases in the two varieties of English in general possible – for example, the recently developed Big Web Corpus (mentioned in the section 2.2.4). Comparing the New Model Corpus (2.2.4) with the BNC could also bring interesting insights into the development of certain idioms over time, similar to those found in Chapter 3.12.

Although I worked with dictionaries from leading publishers, most of them published less than 15 years ago, and some even much more recent, a startling number
of the phrases that I noted down from the dictionaries was not present in any of the three corpora: for example, *yellow as a guinea, Lombard street to China orange, or grow grey in the service of something*. Therefore, I believe that exploiting multi-billion word corpora, like the BiWeC and the NMC mentioned above, might also be beneficial for lexicographers, who could deal with the issue of outdated dictionaries of idioms and related textbooks.
5 Bibliography

5.1 Primary sources


5.2 Secondary sources


“Other services and resources.” British National Corpus. Jan 2009. 5 Feb 2010

<http://www.natcorp.ox.ac.uk/using/index.xml.ID=other#body.1_div.4_div.1>.


“What is the BNC?” British National Corpus. Jan 2009. 5 Feb 2010

<http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=intro>.

“Yellow journalism.” Encyclopaedia Britannica. 2010. Encyclopaedia Britannica


Cambridge Dictionary Online. Cambridge University Press. 3 April 2010

<http://dictionary.cambridge.org/>.


Davies, Mark. BYU-BNC: The British National Corpus. 3 March 2010

<http://corpus.byu.edu/bnc>.


Tkatchuk, Tetyana. “Idioms Inside Out: Money in English Idioms”.


Vokáčová, Jana. “A Comparative View of English, Czech, French and German Idioms”.

APPENDIX: GLOSSARY OF THE CHOSEN IDIOMS OF COLOUR